UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Investigating Semantic Similarity for Biomedical Ontology Alignment

Isabela Mott

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Dissertação orientada por:
Prof. Doutora Cátia Pesquita
João Ferreira

2018

Resumo

A heterogeneidade dos dados biomédicos e o crescimento exponencial da informação dentro desse domínio tem levado à utilização de ontologias, que codificam o conhecimento de forma computacionalmente tratável. O desenvolvimento de uma ontologia decorre, em geral, com base nos requisitos da equipa que a desenvolve, podendo levar à criação de ontologias diferentes e potencialmente incompatíveis por várias equipas de investigação. Isto implica que as várias ontologias existentes para codificar conhecimento biomédico possam, entre elas, sofrer de heterogeneidade: mesmo quando o domínio por elas codificado é idêntico, os conceitos podem ser representados de formas diferentes, com diferente especificidade e/ou granularidade. Para minimizar estas diferenças e criar representações mais *standard* e aceites pela comunidade, foram desenvolvidos algoritmos (*matchers*) que encontrassem pontes de conhecimento (*mappings*) entre as ontologias de forma a alinharem-nas.

O tipo de algoritmos mais utilizados no Alinhamento de Ontologias (AO) são os que utilizam a informação léxica (isto é, os nomes, sinónimos e descrições dos conceitos) para calcular as semelhanças entre os conceitos a serem mapeados. Uma abordagem complementar a esses algoritmos é a utilização de Background Knowledge (BK) como forma de aumentar o número de sinónimos usados e assim aumentar a cobertura do alinhamento produzido. Uma alternativa aos algoritmos léxicos são os algoritmos estruturais que partem do pressuposto que as ontologias foram desenvolvidas com pontos de vista semelhantes – realidade pouco comum.

Surge então o tema desta dissertação onde toma-se partido da Semelhança Semântica (SS) para o desenvolvimento de novos algoritmos

de AO. É de salientar que até ao momento a utilização de SS no Alinhamento de Ontologias é cingida à verificação de *mappings* e não à sua procura.

Esta dissertação apresenta o desenvolvimento, implementação e avaliação de dois algoritmos que utilizam SS, ambos usados como forma de estender alinhamentos produzidos previamente, um para encontrar *mappings* de equivalências e o outro de subsunção (onde um conceito de uma ontologia é mapeado como sendo descendente do conceito proveniente de outra ontologia). Os algoritmos propostos foram implementados no AML que é um sistema topo de gama em Alinhamento de Ontologias. O algoritmo de equivalência demonstrou uma melhoria de até 0.2% em termos de F-measure em comparação com o alinhamento âncora utilizado; e um aumento de até 11.3% quando comparado a outro sistema topo de gama (LogMapLt) que não utiliza BK. É importante referir que, dentro do espaço de procura do algoritmo o *Recall* variou entre 66.7% e 100%. Já o algoritmo de subsunção apresentou precisão entre 75.9% e 95% (avaliado manualmente).

Palavras Chave: Alinhamento de Ontologias, semelhança semântica, algoritmo de equivalência, algoritmo de subsunção

Abstract

The heterogeneity of biomedical data and the exponential growth of the information within this domain has led to the usage of ontologies, which encode knowledge in a computationally tractable way. Usually, the ontology's development is based on the requirements of the research team, which means that ontologies of the same domain can be different and potentially incompatible among several research teams. This fact implies that the various existing ontologies encoding biomedical knowledge can, among them, suffer from heterogeneity: even when the encoded domain is identical, the concepts may be represented in different ways, with different specificity and/or granularity. To minimize these differences and to create representations that are more standard and accepted by the community, algorithms (known as *matchers*) were developed to search for bridges of knowledge (known as *mappings*) between the ontologies, in order to align them.

The most commonly used type of matchers in Ontology Matching (OM) are the ones taking advantage of the lexical information (names, synonyms and textual description of the concepts) to calculate the similarities between the concepts to be mapped. A complementary approach to those algorithms is the usage of Background Knowledge (BK) as a way to increase the number of synonyms used, and further increase of the coverage of the produced alignment. An alternative to lexical algorithms are the structural ones which assume that the ontologies were developed with similar points of view - an unusual reality.

The theme of this dissertation is to take advantage of Semantic Similarity (SS) for the development of new OM algorithms. It is important to emphasize that the use of SS in Ontology Alignment has, until now, been limited to the verification of *mappings* and not to its search.

This dissertation presents the development, implementation, and evaluation of two algorithms that use SS. Both algorithms were used to extend previously produced alignments, one to search for equivalence and the other for subsumption mappings (where a concept of an ontology is mapped as descendant from a concept from another ontology). The proposed algorithms were implemented in AML, which is a top performing system in Ontology Matching.

The equivalence algorithm showed an improvement in F-measure up to 0.2% when compared to the anchor alignment; and an increase of up to 11.3% when compared to another high-end system (LogMapLt) which lacks the usage of BK. It is important to note that, within the search space of the algorithm, the *Recall* ranged from 66.7% to 100%. On the other hand, the subsumption algorithm presented an accuracy between 75.9% and 95% (manually evaluated).

Keywords: ontology matching, semantic similarity, equivalence algorithms, subsumption algorithms

Resumo Alargado

Nas últimas décadas o domínio biomédico tem tido uma explosão de informação com o registo de vários estudos e técnicas, e.g. sequenciação de DNA. Essa informação muitas vezes é publicada em linguagem natural dificultando o tratamento computacional; outras vezes é armazenada em bases de dados que, mesmo dentro do mesmo domínio, podem ser desenvolvidas de forma distinta complicando a partilha de informação, e posterior extração de conhecimento. As ontologias têm sido muito utilizadas neste domínio não só pela capacidade de lidar com a heterogeneidade dos dados, mas por facilitar a interoperabilidade entre máquinas, entre humanos, e entre máquinas e humanos.

As ontologias são conjuntos de conceitos ligados entre si por relações de forma a descreverem um domínio. Entre outras características, os conceitos podem ser associados a descrições, propriedades e sinónimos. Ao nível das relações estabelecidas, as mais comuns são as de subsunção (em inglês is_a), relações que representam uma hierarquia simples onde o descendente herda todas as características do seu ancestral e torna-se mais específico ao ter propriedades próprias. No entanto, devido à tendência para criar novas ontologias por parte de equipas de investigação, observa-se o aumento da heterogeneidade entre as próprias ontologias, pois investigadores com problemas diferentes e diferentes pontos de vista podem gerar ontologias com representações da realidade diferentes. Neste contexto, existem três tipos de heterogeneidade: a heterogeneidade de domínio que ocorre quando uma ontologia descreve um domínio ou subdomínio distinto de uma segunda ontologia; no caso de duas ontologias descreverem o mesmo domínio ou domínios semelhantes pode existir uma heterogeneidade a nível do detalhe utilizado por cada uma delas; ou então uma heterogeneidade ao nível da interpretação utilizada durante o desenvolvimento.

Para lidar com esta heterogeneidade existem algoritmos (*matchers*) que procuram alinhar as ontologias, isto é, encontrar conceitos de duas ontologias distintas que são equivalentes ou estão relacionados de outra forma (*mappings*). Um conjunto de *mappings* é denominado de Alinhamento. Um *mapping* é constituído por um conceito de uma ontologia, outro conceito de uma segunda ontologia, a relação entre eles e um valor de semelhança. Os *mappings* mais comuns são os de equivalência; outro tipo de *mappings* são os de subsunção onde um conceito é mapeado como sendo descendente do outro.

Dentro dos algoritmos de alinhamento, o tipo mais comum é aquele em que o valor de semelhança é calculado através da informação léxica dos conceitos. Muitas vezes são desenvolvidos algoritmos rápidos que produzam um alinhamento âncora que possa ser estendido por algoritmos mais complexos, por exemplo algoritmos estruturais, que usam a informação estrutural das ontologias para estender um alinhamento âncora. Os algoritmos estruturais partem do pressuposto de que as ontologias têm uma interpretação semelhante, o que nem sempre é verdade. Assim, o âmbito desta dissertação é a implementação de algoritmos de alinhamento de ontologias baseados em Semelhança Semântica (SS) como alternativa aos algoritmos estruturais. Dentro da literatura a SS é apenas utilizada como uma forma de validar os *mappings* encontrados por outros algoritmos.

A Semelhança Semântica devolve um valor numérico que reflete a semelhança entre dois conceitos dentro de uma ontologia, valor que é obtido tendo em conta a estrutura da ontologia e o significado dos conceitos comparados. A cada conceito é atribuído um valor de conteúdo de informação (IC - do inglês *Information Content*) que é usado posteriormente no cálculo da distância semântica entre os conceitos através de medidas de SS.

De forma a estender um alinhamento âncora foram desenvolvidos dois novos algoritmos utilizando SS, um deles com o objetivo de encontrar

mappings de equivalência e outro *mappings* de subsunção. Os *mappings* existentes nesta âncora são utilizados como ponto de partida para encontrar *mappings* candidatos na sua vizinhança.

O algoritmo de equivalência necessita de um segundo alinhamento de *input* com um *threshold* inferior ao do alinhamento âncora. Os *mappings* do alinhamento com *threshold* inferior presentes na vizinhança de um *mappings* candidato irão ser utilizados juntamente com a SS dos conceitos dentro de cada ontologia para computar uma contribuição semântica. O valor de semelhança final do par candidato terá em conta as contribuições semânticas da sua vizinhança. O algoritmo criado para encontrar equivalências depende de vários parâmetros que podem ser agrupados em três dimensões: parâmetros que definem como é que o algoritmo explora a **estrutura** das ontologias a serem alinhadas, parâmetros que definem como **calcular semelhança semântica**, e parâmetros que definem como **pesar** os valores de semelhança semântica de forma a incrementar a pontuação de um *mapping* candidato. O algoritmo foi avaliado ao nível destas três dimensões utilizando uma versão simplificada do alinhamento gerado pelo AgreementMakerLight (AML). É importante referir que os algoritmos de alinhamento criados são independentes do sistema em que estão incluídos, no entanto a sua implementação foi feita no AML, um sistema de Alinhamento de Ontologias topo de gama para Ontologias Biomédicas e também porque ser um sistema modular onde facilmente se implementam novos *matchers*.

Após serem feitos testes preliminares com o alinhamento simplificado do AML com todas as combinações possíveis dos parâmetros de cada dimensão, o alinhamento produzido pelo AML completo foi utilizado como âncora. Para o AML completo não foram registados aumentos significativos na performance do algoritmo de alinhamento. Contudo, ao utilizar o alinhamento simplificado como *input*, foi possível aumentar valores de *F-measure* em 0.2%. Estes resultados demonstram que o alinhamento de input poderá ter influência no espaço de procura. O

Recall dentro do espaço de procura do algoritmo variou entre 66.7% e 100%, o que indica que de facto o novo algoritmo de alinhamento foi capaz de encontrar uma quantidade substancial de novos *mappings* dentro do raio de procura sem diminuir de forma significativa a precisão. Dado o facto de que o algoritmo apresenta valores de *Recall* elevados dentro do espaço de procura, foi feita uma comparação com outros sistemas de Alinhamento de Ontologias (LogMapLt e LogMapLt). Ao comparar os resultados gerados com o alinhamento simplificado com os resultados do LogMapLt foi observado um aumento de *F-measure* de 11.3%.

Relativamente ao algoritmo de subsunção, este também depende de *mappings* candidatos. Neste algoritmo os *mappings* candidatos também incluem os irmãos dos *mappings* âncora. Os irmãos podem ser considerados descendentes ou ancestrais do outro conceito envolvido no *mapping*. Cada um dos *mappings* candidatos tem o seu nome e sinónimos normalizados e um novo valor de semelhança é computado com base nesta normalização. Para a avaliação dos novos *mappings* foi feita uma avaliação manual de 332 novos *mappings* finais escolhidos aleatoriamente, recorrendo aos sinónimos e definições disponíveis nas próprias ontologias, a dois dicionários médicos e a artigos científicos. A avaliação categorizou os novos *mappings* em: subsunção correcta, subsunção com direção contrária à esperada, equivalência e incorrecto. A precisão destes *mappings* assume apenas como corretos os que foram considerados como "subsunção correta", sendo que os valores variaram entre 75.9% e 95%.

Outras contribuições prestadas durante a duração desta dissertação foram a apresentação de um poster *"Integrating semantic distances in ontology matching algorithms for biomedical ontologies"* no *Biomedical Open Days*; a participação no *Ontology Alignment Evaluation Initiative* na tarefa de *Disease and Phenotype* pela equipa do AML; e coautoria de um artigo publicado no *Journal of Biomedical Semantics*, no âmbito de uma pesquisa relativamente à capacidade dos sistemas

de Alinhamentos de Ontologias de lidarem com os problemas inerentes às ontologias Biomédicas (Faria *et al.*, 2018).

Os resultados obtidos não rejeitam a hipótese de que a Semelhança Semântica possa ser utilizada para estender alinhamentos existentes. Em suma, a utilização de Semelhança Semântica para a extensão de alinhamentos âncora com uma alta precisão pode ser uma opção válida para o Alinhamento de Ontologias de domínios onde: i) o *Background Knowledge* está indisponível ou é difícil de ser explorado; ii) ou em situações do mundo real em que os altos níveis de otimização obtidos pelos sistemas atuais de Alinhamento de Ontologias não sejam viáveis (por exemplo, os atuais sistemas não foram desenvolvidos tendo em vista o domínio das ontologias financeiras). Esta dissertação poderá ser continuada ao nível de implementação dos algoritmos em outros sistemas de Alinhamento de Ontologias; e ao nível de testes com pares de ontologias provenientes de domínios sem BK ou em situações reais onde os altos níveis de otimização, por parte dos sistemas existentes, não sejam possíveis.

Acknowledgements

I would like to start by thanking my advisor Professor Cátia Pesquita and my co-adviser Professor João Ferreira for giving me the right tools, guidance, and opportunity throughout this dissertation. The work of this dissertation was fully finded by the Fundação para a Ciência e Tecnologia through the project SIMILAX (PTDC/EEI-ESS/4633/2014).

There is a set of people that have also helped without whom this dissertation would have not been completed by this time. These people have helped me through the course of this work and/or the correction of this dissertation, they are Mafalda, Joana, Fernando, Vinicius, and Soraia. I also have to thank the people that were and are a support during my Masters: Pedro, Gonçalo, and Vera, my Masters colleagues.

Last but not least my family: mother, both sisters, my aunt, uncle, and Dobby. To whom I need to thank for all the patience, as well as apologize for all the events missed due to this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For the past decades, research in the biomedical field has been generating a significantly large amount of data through several biological studies and techniques, such as DNA sequencing, genome annotation, image analysis, chromosome topology, protein localization, or clinical data. Biomedical data is often recorded in natural languages. For instance, scientific publications and clinical annotations resulting from a continuously growth of the available data and databases for researchers. However, the extraction of knowledge from the data remains a big challenge. The usage of databases alone does not entirely solve the problem concerning the organization of information, and different biomedical databases may not be interoperable.

In order to satisfy the previous needs, the biomedical field has been using ontologies, which describe concepts related to a domain of knowledge, their properties, and the relationships between them. When biomedical data in different databases is described under a common model provided by an ontology, it becomes interoperable. Thus ontologies help solve the issues of heterogeneity in biomedical databases.

Ontology development and usage has gained acceptance in this particular field not only for its ability to correlate concepts (e.g., one concept descends from another, resembling the idea of taxonomy), but also given its resemblance to a dictionary (i.e., an ontology contains annotations and properties such as synonyms). These characteristics make ontologies quite useful in cases where we are dealing with the intrinsic ambiguity of this field, but also with the exponential growth of

information, allowing the creation of ontologies for several smaller domains (e.g., anatomy or chemistry), and even for sub-domains or sub-disciplines (e.g., human or mouse anatomy).

Both the size and specificity of a domain can be transposed to an ontology. Two ontologies can be specially developed to suit domains that can be as distinct as anatomy and ecology or, as similar as mouse anatomy and human anatomy. Ontologies with distinct domains complement each other in a macroscopic level and can encode, for example the whole biomedical domain, whereas different ontologies with similar or even equal domains can encode, for the same concept, different perspectives or relationships. In fact, ontologies covering the same domain have been independently developed by different groups, creating heterogeneity at the ontology level. For example, in the Mouse Adult Gross Anatomy Ontology the term "Female germ cell" is described as part of the "Female reproductive system" having "Oocyte" as both synonym and related synonym (Hayamizu *et al.*, 2005); whereas in the Foundation Model of Anatomy, the term "Oocyte" is described as "Germ cell of the female sex" (Rosse & Mejino, 2003).

Ontology matching has been used to solve these heterogeneity problems by finding bridges of knowledge between two concepts from two different ontologies - mappings. A mapping links two equivalent concepts (equivalence mapping) or concepts hierarchically related (subsumption mappings), whereas the concept from one ontology is mapped as more generic (ancestors) than the concepts from the other ontology which are more specific (descendants). By matching similar concepts, ontology matching algorithms allow machines to more easily integrate their knowledge by leveraging the existing synonyms (i.e., concepts that share properties and annotations), and compare ancestors and/or descendants. Those aspects allow possible gaps to be filled in, or even information complementation.

The idea of linking those entities (i.e., connecting the ontologies) helps to further increase interoperability between different data sources, as well as enhance the existing knowledge.

1.1 Motivation and Objectives

The biomedical domain is vast and ambiguous, which makes it a challenge for ontology matching. Most of the existing ontology matching systems rely on a lexical approaches, i.e., they explore the ontology vocabulary to find mappings.

Since biomedical ontologies are typically very large, some ontology matching systems use fast lexical algorithms to produce a set of mappings (alignment) that serve as an anchor (anchor alignment) to be extended by using more complex algorithms. This extension is based on the ontology's structure and defined by the path distance between ontology classes. However, distance in an ontology graph is a poor approximation of semantic distance, since it is common that links in different areas of the ontology or at different depths convey different semantic distances.

The main goal of this dissertation is to create novel algorithms that explore semantic similarity as an extension technique for ontology matching. Semantic similarity metrics that give a numeric score to the closeness in meaning between two concepts in the same ontology. Exploring this notion, lexically derived mappings can be used as starting points to derive new mappings based on semantic similarity. The underlying hypothesis is that, using semantic similarity to an established mapping to increase a candidate mapping score, is a valid approach for alignment extension.

This approach will be investigated in two areas: (i) the development of algorithms for equivalence mappings; and (ii) the development of algorithms for subsumption mappings. Most of the state of the art algorithms search uniquely for equivalence mappings. By creating algorithms for subsumption, the interoperability of ontologies is further increased, particularly in the cases where ontologies possesses different levels of granularity of modeling details. This is particularly important for cases where an ontology for a specific domain (with high granular knowledge of a specific part of that domain) is being mapped to a more general purpose ontology.

1.2 Contributions

The specific contributions of this dissertation can be enumerated as follows:

1. Creation of one algorithm of ontology matching that relies on semantic similarity measures to create equivalence mappings.
2. Creation of another algorithm of ontology matching that also relies on semantic similarity measures but this time to create subsumption mappings.
3. Manual evaluation of subsumption mappings in the context of this dissertation.
4. Poster in the Biomedical Open Days titled "Integrating semantic distances in ontology matching algorithms for biomedical ontologies".
5. Participation in OAEI 2017 in the Disease and Phenotype track as part of the AML team (**?**).
6. A survey of state of the art ontology matching systems regarding their ability to address the specific challenges in biomedical ontology matching was conducted. Submission under revision of an invited article for the Journal of Biomedical Semantics (Faria *et al.*, 2018).

1.3 Overview

This first Chapter, Introduction, serves as a presentation of this dissertation including its motivation. In Chapter 2, Concepts and Related Work the concepts needed for this dissertation are presented and explained as well as the relevant work done in this field to this date. Methods is the third Chapter, where the proposed algorithms are presented; their Evaluation Methodology is presented in Chapter 4. The Results and Discussion are stated in Chapter 5, and the Conclusion is in Chapter 6.

Chapter 2

Concepts and Related Work

This chapter introduces the basic concepts required to understand the presented work, namely ontologies, ontology matching, and semantic similarity. It also describes relevant work in ontology matching literature.

2.1 Ontology

The term *Ontology* was re-purposed by Gruber in early 1990's, for the context of computer science, to mean "an explicit specification of a conceptualization" (Gruber, 1995). Gruber (2008) has updated this definition, for the same context in 2008, by defining an ontology as a set of primitives (explained further in this chapter) modeling a domain of knowledge.

While information has been exponentially increasing, new information is created, and obsolete and/or incorrect information is upgraded. Humans' ability to follow this increase in terms of knowledge has been insufficient. As a response to the problem identified in the introduction, knowledge bases have been used not only for data but also knowledge storage - such as ontologies. The application of ontologies on various fields has been a major aid. For example, ontologies can be very helpful for researchers to process data by being a sharing mechanism not only between humans, but also as a bridge that connects humans and machines.

Usually, an ontology defines a vocabulary used by a particular domain. In other words, ontologies enable the organization of concepts that can be used to

describe domains while allowing computer reasoning, which is the result of logic rules known as axioms, and which generate new knowledge.

Regarding the referred primitives, those are typically classes, attributes, and relationships. Ontologies often structure their concepts and the relationships between them as a Directed Acyclic Graph (DAG), meaning entities are nodes and relationships are edges, as illustrated in Figure 2.1. There are many different types of entities. Main entities are *classes*, or terms, representing a set of individuals from that domain. *Instances* correspond to objects, particular individuals of a domain. *Relations* represent the existing links between concepts. *Data-types* specify value types (e.g., data-type String), and finally *data values* are values stated in agreement with the data-type (e.g., the data-type of the "label" property is String, and a possible value for this property is "Cell").

In Figure 2.2, both arm and leg are specifications of limb, being synonyms to upper and lower limbs respectively. An *is_ a* relationship specifies that a descendant will inherit all the properties from its ancestors plus its own specifications. The *part_ of* relationship implies that the descendant may only exist as a part of the ancestor but the ancestor might exist regardless of the descendant.

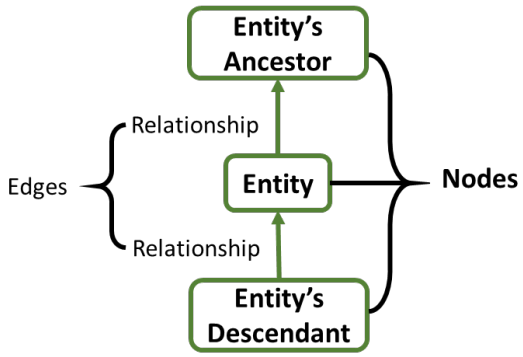


Figure 2.1: Schema of ontologies basic component organization

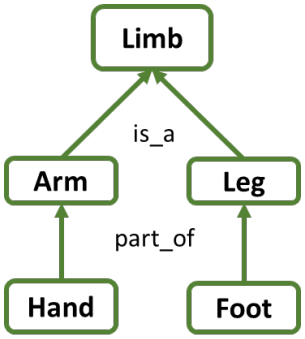


Figure 2.2: Example of a possible small fragment from a biomedical ontology

One of the main reasons for the development of ontologies is their ability to function in multiple roles (Euzenat *et al.*, 2007):

- Share a common understanding of the knowledge about a domain among people or software agents.

- Reuse of domain knowledge, where this knowledge can be shared. For example, concepts that are described in a generic ontology could be used as a starting point for a more specific one. A second example can be the case where one ontology expands into a different domain borrowing concepts from a second one (more focused and complete in the latter specific domain).

- Make domain assumptions explicit, meaning there are no hidden assumptions. Even concepts that are generic and/or shared are explicit in order to avoid ambiguity. For example, the concept *"Cell"* is very common in this domain but has different descriptions depending on the approach, i.e., this concept might represent the basic unity of all organisms or an anatomical structure depending on the point of view of the ontology.

- Separation of domain knowledge from the operational knowledge. Domain knowledge derives from the domain which can be described into different ontologies, while operational knowledge is independent of the domain. For example, an algorithm that was developed for a certain domain handles domain knowledge, while the algorithm itself is operational knowledge.

- Analysis of domain knowledge since an ontology represents a domain (or its part). Therefore the content analysis can be representative of that domain (or its part).

Different types of ontologies can be distinguished by their level of formality. The smallest level of formality is a simple list of concepts while higher levels can be achieved by adding relationships between concepts. This incrementation of complexity can result from the usage of synonyms/antonyms or more complex relationships, such as meronyms (relationship between the whole and a specific part, i.e., the relationship stated with the *part_of*).

There are two notions that have not been yet presented but need clarification for an accurate understatement: leaves and depth. In the context of ontologies, a leaf is a concept that has no descendants. The depth of a concept is the distance between that concept and the root. Thus the Maximum Depth is the depth of the whole ontology.

2.1.1 Biomedical Ontologies

The biomedical field has been taking advantage of ontologies for the past decades. Creation and development of *Biomedical Ontologies* can aid in many intrinsic problems of this field (e.g., ambiguity of natural language). When biomedical information is stored in an ontology it allows consensus through a group of synonyms instead of an unique word. Decision support in health care not only presents the challenges stated but also represents a good example of the application of ontologies to this area. It is important to emphasize that the previous has been the reason behind the existence of *Biomedical Ontologies*.

There are many successful biomedical ontologies, such as Human Phenotype Ontology (HP) (Köhler *et al.*, 2013) or Gene Ontology (GO) (Ashburner *et al.*, 2000), which are able to handle not only the vast amount of knowledge that result from this field, but also the heterogeneity that is a result of the biological and clinical data. The development of tools and techniques to explore ontologies that belong to the biomedical domain must consider the specific characteristics of these ontologies (Faria *et al.*, 2018):

- **Large Size:** biomedical ontologies allocate thousands of classes representing a computational challenge.
- **Complex Vocabulary:** besides the natural text ambiguity this particular domain encodes several names for the same class, and for the same main label there can even be different kinds of synonyms (e.g., narrow or broad synonyms).
- **Rich Axioms:** semantics has been one of the subjects taken into consideration within this topic, since axioms can vary from the simplest *is_a* (e.g., an eukaryotic cell *is_a* cell) to *part_of* (e.g., the nucleous is *part_of* the eukaryotic cell), to more complex ones. For example, a possible more complex axiom could be "a person has a child", and since a child *is_a* person and a person is either boy or a girl, the machine is be able to assess that that person has either, at least, one boy or one girl.

Biomedical Ontologies in general have common features such as: stable and unique codes for biomedical concepts, preferred labels, synonyms, external sources of information, various textual descriptions and cross-links, and frequent updates by domain experts. This vast domain has the ability to address an array of problems similar in size, such as the search and query of heterogeneous data, data exchange among applications, information integration, natural language processing, representation of knowledge, computer reasoning with data, and information retrieval (Rubin *et al.*, 2007). The usage of biomedical ontologies as tools for data integration and annotation can provide a common and needed vocabulary. The latter can improve description, communication of results, and the creation of bioinformatic tools for analysis of microarray data, network modeling, and so on.

Bioportal is the largest repository of Biomedical Ontologies with over 500 ontologies. It is important to mention this Open Source tool since it will be referred through this paper but detailed information falls over the scope of this dissertation. Therefore, for more information on Bioportal see Whetzel *et al.* (2011).

2.2 Semantic Similarity

From a syntactical point of view the words *arm* and *leg* are not similar, but when presented as in Figure 2.2 a machine is able to read their Ancestors as siblings, and understand both are siblings. When associating concepts to a domain, e.g., anatomy, terms are linked by relations, making their relationships readable by a machine.

Semantic similarity is the computation of the similarity between entities based on their meaning, as described in an ontology. As stated, comparing entities can be facilitated by objective representations and measurable properties while the difficulty is increased by functional aspects which have no direct comparison (Pesquita *et al.*, 2009). The latter stimulates ontologies annotation for further comparison, as well as various types of matchers. Comparing terms using semantic similarity can be divided in two main categories of measures, that are distinguished by the type of data-structure used: edge and node based measures.

Edge-based Approaches consist mainly in counting the edges between terms. The most trivial approach returns the distance between the two terms; when in the presence of more than one path the distance is presented either by the shortest or the average route. Even though this type of approach is intuitive and easily converted into a measure, their foundation lies on equal distribution of terms through the ontology. Equality in distribution implies that nodes and edges have an uniform structure, as well as each level of an ontology sharing the same granularity. Those are both uncommon scenarios in biological ontologies (Alexander, 2006). Despite of the effort to deal with those issues, the current strategies only attenuate instead of solving them (e.g., by normalizing the edges based on their hierarchical depth maintains equal values for edges at the same hierarchical level).

Node-based Approaches were created by changing the focus to nodes instead of edges. Resnik (1995) proposed to take into account the corpus involved. A frequent notion in this branch is the information content (IC) which is a numeric value representing the specificity of the information that is enclosed inside an entity. The closer a term is to the ontology root, the more generic it is, and less specific information it carries, leading to a lower IC. Since the idea is to compute the semantic similarity between two terms, the IC by itself is not enough considering it is generated for one entity. There are various ways not only to compute ICs but also to calculate semantic similarity.

The three approaches to compute IC used in this dissertation were:

- $IC_{Seco}(c)$ (Seco *et al.*, 2004) refers to the complement of the division between the direct and indirect descendants from c (including c) - $N_d(c)$, and the total number of concepts in the ontology - N (equation 2.1).

$$IC_{Seco}(c) = 1 - \frac{log(N_d(c))}{log(N)} \tag{2.1}$$

- $IC_{Zhou}(c)$ (Meng *et al.*, 2012) takes a parameter k as the contribution of each estimator. The first one is the $IC_{Seco}(c)$, and the second the division between c's depth and the ontology's (equation 2.2). K was defined by the

external library used to compute the semantic similarity.

$$IC_{Zhou}(c) = k*(IC_{Seco}(c))+(1-k)\frac{log(max(depth(c))}{log(depth_{max})}, \text{ in this work k} = 0.5 \tag{2.2}$$

- $IC_{Sanchez}(c)$ (Sánchez *et al.*, 2011) considers the percentage of $leaves(c)$ (number of leaves in the descendants of c) in the descendants and the concept itself as well as the whole number of existing leaves in the ontology (equation 2.3).

$$IC_{Sanchez}(c) = -log\frac{\frac{leaves(c)}{descendants+1}+1}{max_{leaves}+1} \tag{2.3}$$

When using ICs to compute Semantic Similarity, the measures used rely on a technique based on common ancestors: MICA - Most Informative Common Ancestor, meaning the common ancestor of the two concepts being compared with highest IC. All the measures will be presented as the computation of the semantic similarity between the terms u and v - $sim(u, v)$:

- Starting with the simplest measure, Resnik (1995) which is just the IC from the first common Ancestor (equation 2.4).

$$sim_{Resnik}(u, v) = IC(MICA_{u,v}) \tag{2.4}$$

- Lin's measure uses the $MICA_{u,v}$ weighted by the IC from both terms (equation 2.5).

$$sim_{Lin}(u, v) = \frac{2 * IC(MICA_{u,v})}{IC(u) + IC(v)} \tag{2.5}$$

- Jiang and Conrath approach is nothing more than a distance, the addition of 1 is a normalization - equation 2.6.

$$sim_{JC}(u, v) = \frac{1}{IC(u) + IC(v) - 2 * IC(MICA_{u,v}) + 1} \tag{2.6}$$

- Finally, GIC's measure (Pesquita *et al.*, 2008) where the measure is given by the sum of each term's IC in the intersection of u with v divided by the

sum of their union (equation 2.7).

$$sim_{GIC}(u,v) = \frac{\sum IC(t)}{\sum IC(w)},\ t \in (u_{Anc} \cap v_{Anc}) \text{ and } w \in (u_{Anc} \cup v_{Anc}) \quad (2.7)$$

2.3 Ontology Matching

One of the obstacles that urged the development of ontologies was the heterogeneity of data (e.g., synonyms). By allowing a main label to be related to other labels through relations (e.g., synonyms) ontologies minimize the ambiguity resulting from natural language. A new problem arises from this solution: heterogeneity of ontologies.

When dealing with ontologies there are four main types of heterogeneity that are relevant (Euzenat *et al.*, 2007): *Syntactic heterogeneity* resulting from two ontologies expressed in different languages; *Terminological heterogeneity* refers to the synonym problem meaning that two equal concepts might have different labels in two ontologies; *Conceptual heterogeneity* is the modeling discrepancy, meaning that one domain might be modeled from two points of view, covering different sub domains, or having different levels of detail; *Semiotic heterogeneity* relates to the bridge between humans and machines. While humans easily distinguish homonyms by their context machines do not. This heterogeneity will not be discussed any further for lack of relatedness with this work. Thus, the logical definitions for the presented concepts are as follows (Euzenat *et al.*, 2007):

One way to decrease the existing heterogeneity is linking different ontologies - *Ontology Matching*. The aim in Ontology Matching is to create bridges of knowledge. The previous can be achieved through matchers. *Matchers* are algorithms that use different strategies to calculate the similarity between two entities in two different ontologies. When this similarity is associated to those terms, a *Mapping* is created. A set of mappings above a certain value (threshold) that passes through a selector is called an *Alignment*. The difference between a set of mappings and an Alignment lies on the selection step that will deal with cardinality problems.

Ontology matching systems rely on different types of matchers in order to generate alignments. Usually when computing alignments, the matchers try to

find equivalence mappings - two entities from different ontologies that represent the same concept. Alternatively, there are *subsumption* mappings, i.e., mappings from one ontology to another one that convey the meaning of a hierarchical relationship between the two concepts. This idea is easily pictured when dealing with an horizontal ontology lacking structural detail but excelling on the quantity of concepts.

It is important to emphasize that subsumption mapping means one term is more generic that the other. The idea behind subsumption mappings is to deal with both perspective and conceptual heterogeneities by being able to complement ontologies with, for example, missing detail levels. By finding the subsumption matches within a pair of ontologies, the structure of a vertical ontology and the width of an horizontal one can be used to better organize the existing domain knowledge.

When creating ontologies there are some types of heterogeneity that might be involved. When dealing with ontologies that describe one domain but different points of view there can be some adjacent problems such as different levels of detail. If the previous problem did not exist there would be no need for *Subsumption Matchers*, since the integration of ontologies would automatically create the same output as the *Subsumption Matchers*. There is a large diversity of information in the pursuit to gather knowledge. For example, there are ontologies that while describing the same domain might consider C as A's ancestor ($C = A_A$) while others recognize C as a sibling of A. Euzenat *et al.* (2007) classifies ontology matching techniques into two categories:

1. *Granularity/Input Interpretation*, where the classification is based on the matcher's granularity (element or structure level) and then on how the input's information is interpreted:

    - Element-level matching techniques compute mappings by analyzing the concepts in isolation, i.e., ignoring their relations (Rahm & Bernstein, 2001).
    - Structure-level techniques compute mappings by taking into account the concept's surrounding relations (Kang & Naughton, 2003).

2. *Kind of Input*, based on the usage done by the matching techniques to the kind of input.

    - Syntactic techniques interpret the input regarding its structure following an algorithm.
    - External techniques uses auxiliary resources of a domain and common knowledge external to the algorithm. Semantic techniques take advantage of formal semantics.

2.4 Related Work

Most of the existing matchers find equivalences by lexical or string similarity. Their scoring relies on the similarity either of the words composing the labels or the synonyms.

The AgreementmakerLight (AML) is a top performing system that will be used as a base for the implementation of the methods used through this dissertation. In terms of matchers it has word-based string similarity matchers using synonyms based on queries to the WordNet database. AML also implements a variety of similarity metrics and weights to compute string similarity, requiring an all-against-all comparison of words. Since this system will be further discussed in section 4.1.2, at this point is important to say that there is no usage of semantic similarity measures in AML.

LogMap is a "highly scalable ontology matching system with built-in reasoning and diagnosis capabilities" (Jiménez-Ruiz *et al.*, 2011). Nowadays LogMap is a family of systems that includes the complete LogMap, LogMapLt, and LogMap-Bio (Jiménez-Ruiz *et al.*, 2016). LogMap's complete pipeline can be summarized by a first overlapping estimation with a lexical algorithm that over estimates the potential mappings, followed by a lexical indexation where not only the labels but also their lexical variations are indexed. The possible mappings are then separated into two groups by an heuristic that divides them into correct mappings or in need of expert curation. At this point there is a mapping repair and a structural indexation. The previous step allows conflict detection using disjointness axioms. Finally LogMap allows user intervention whereas an human expert

will approve or reject the mappings. LogMapBio is an extension of LogMap by including the usage of Bioportal to provide mediating ontologies. LogMapLt is a variant of LogMap that reduces the latter to only its string matching techniques.

Is important to emphasize that Cross & Hu (2011) reviewed the literature involving the usage of semantic similarities in ontology matching. Unfortunately, the gap in years resulted in the fact that the relevant analyzed systems do not appear in recent literature. Systems such as ASMOV (Automated Semantic Matching of Ontologies with Verification), last updated in 2010 reportedly utilized semantic similarity to verify the mappings found with lexical and string similarity algorithms. For more information, see Jean-Mary *et al.* (2009). For the best of my knowledge there are no current systems that use semantic similarity to compute new mappings. The advantage taken of semantic similarity falls under the verification mechanism's scope like ASMOV.

Even though there has been a bigger investment towards equivalence mappings, subsumtion has always been an ongoing area of research. Competition wise, there has been a renewed interest in this type of mappings. Usually, as in AML, the system creates equivalence matchers and reuses them to find subsumption mappings by making small changes. Nowadays the only system that has presented subsumption matching as a priority is PhenomeNET (Garcıa *et al.*, 2016). PhenomeNET contains classes from multiple ontologies in order to preprocess disjoint axioms. The main goal is to integrate specific ontologies having species phenotypes based on Entity-Quality (EQ) definition patterns. EQ will gather entities from one ontology and increase their quality with another. This ontology uses entities from UBERON (cross-species ontology for anatomic structures) and quality information from PATO's (phenotypic qualities/properties) ontology (Mungall *et al.*, 2012). Therefore their matchers have a specific affinity for ontology pairs that are annotated with PATO, leaving other pairs in this domain short ended. Also, ontologies that do not fall into the biomedical domain will not be matched (Jiménez-Ruiz *et al.*, 2016).

Table 2.1 (Faria *et al.*, 2018) describes each system according to its capabilities: to handle *Size* of the ontologies, medium-sized (+), large (++) or very large (+++) ontologies; use of *Lexicons* thereforethe possibility of synonym usage and the lexical tools such as WordNet (WN) or UMLS SPECIALIST Lexicon

Table 2.1: Ontology matching systems that have participated in OAEI for the biomedical tracks.

System	Size	Lexicon	Relations	Repair	Background Knowledge	OAEI BT
AgrMaker [6]	+	weights	*part of*	-	Bio; Man; Med	A
AML [13]	+++	WN; weights	all	Logic	Bio; Auto; M/E	all
Anchor-Flood [55]	+	WN	-	-	Man; Exp	A
Aroma [7]	+++	-	-	-	-	A, LB
ASMOV [24]	+	WN	-	-	Bio; Man; Exp	A
AUTOMSv2 [31]	++	WN; weights	-	-	Man; Exp	LB-
BLOOMS [42]	+	WN	*part of*	-	Bio; Man; Med	A
COMMAND [36]	+	-	-	Logic	-	A
CroMatcher [19]	+	WN	-	-	Man; Exp	A
DKP-AOM [11]	+	WN	-	-	Man; Exp	A, LB-
DSSim [38]	+	WN	-	-	Man; Exp	A
FCA-Map [64]	++	UMLS	-	Logic	Man; Exp	all-
GMap [33]	+	external	-	Logic	Man; Exp	A
GOMMA [29]	+++	-	-	-	Bio; Auto; Med	A, LB
kosimap [48]	+	-	-	-	-	A
LogMap [26]	+++	WN; UMLS	-	Logic	Bio; Auto; M/E	all
LP HOM [34]	+	-	-	-		A
Lyam++ [59]	++	BabelNet	-	-	Man; Exp	all-
MapPSO [4]	+	-	-	-	-	A
OACAS [63]	+	-	-	-	-	A
PhenomeNET [15]	++	(AML)	*part of*	-	Bio; Man; Med	DP
SAMBO [32]	+	WN	*part of*	-	Bio; Man; Exp	A
ServOMap [8]	+++	WN	-	Logic	Man; Exp	A, LB
TaxoMap [20]	+	-	-	-	-	A
TOAST [58]	+	-	-	-	-	A
WikiMatch [23]	++	Wikipedia	-	-	Man; Exp	A, LB-
YAM++ [39]	+++	WN	-	Rules	Man; Exp	A, LB

Relations lists the types of relations they contemplate in addition to subclass relations; *Repair* details whether they perform alignment repair based on logic or rules; *Background Knowledge* describes whether they use biomedical ontologies as background knowledge (Bio), whether the process of background knowledge selection is manual (Man) or automatic (Auto), and whether background knowledge is used as a mediator (Med) or for lexical expansion (Exp); *OAEI BT* lists the Biomedical Tracks in which the system successfully competed in out of Anatomy (A), Large Biomedical Ontologies (LB), and Disease & Phenotype (DP), with - indicating that the system did not complete the largest LB tasks.

(UMLS); set of *Relations* contemplated by the system besides subclass; type of *Repair* performed as part pf the alignments logic rule; *Background Knowledge* and if background knowledge's selection is manual (Man) or automatic (Auto), plus if it is used as a mediator (Med) or for lexical expansion (Exp);*OAEI Bio Tracks* where the system had completed Anatomy (A), Large Biomedical Ontolo-

gies (LB), Disease & Phenotype (DP) and '-' indicates that the system did not complete the largest LB tasks.

From this overview it becomes clear that only two systems participated in all tracks. Both systems, AML and LogMap, explore Background Knowledge (external information) specific to the biomedical domain.

Chapter 3

Methods

Two kinds of matchers were developed to investigate the application of different semantic similarity approaches in the context of ontology matching: equivalence matchers and subsumption matchers. It is important to emphasize that the matchers created find new mappings by extending an input alignment, which means that they require an anchor alignment where the mappings will function as anchors for the search of new mappings.

Upon matching two ontologies (O_1 and O_2) a matcher will find bridges of knowledge between them. Those bridges, called mappings, are composed of a term A from O_1, a term B from O_2, a relationship r between A and B, and a score that reflects the confidence assigned to the pair. When a matcher focuses on finding mappings that represent the same real-world concept they are called *Equivalence Matchers*. The notion behind *Equivalence Matchers* developed in this work is based on taking that score bellow the alignment threshold but above a second defined threshold and evaluating their semantic similarity to an accepted mapping. The usage of *Semantic Similarity* to find subsumption mappings also relies on a neighborhood search for new mappings.

3.1 Equivalence Matchers

Each matcher in this category needs as input, two alignments and parameters from three Dimensions, as well as a final threshold. A threshold is just a value that the candidate mapping needs to surpass to become a mapping. One alignment

will be the anchor for the extension of the alignment while the second will have a lower threshold ($Alignment_{LT}$) that will be used as a source of initial similarities for an incrementation by semantic similarity computation. This is not a true alignment since a selection procedure is not applied to it.

The anchor alignment will be composed of anchors, each anchor will be the starting point to a search for new mappings on its neighborhood. While a candidate mapping will not be in the anchor alignment but in $Alignment_{LT}$, this type of mapping will also present a score of semantic similarity. Only if the combination of the previous scores surpasses the threshold of the matcher is a mappings considered a true mapping and even then it will only be a part of the final alignment if it passes the selector step.

Given that the proposed method is parameterizable in several ways, the parameters were grouped in 3 distinct dimensions, based on their relation with the algorithm:

- **Structure -** takes into account the direction of the semantic expansion as well as the distance that is used. The direction can be Ancestors (A), Descendants (D), or both. When dealing with both, the direction chosen is the one that contains the Maximum (M) score.
- **Semantic Similarity -** takes into account the IC measure and the Semantic Similarity Measure. The matching algorithm can take as input any IC measure and any semantic similarity measure that function over ontologies.
- **Weighting Mechanism -** is the way the semantic similarity is weighted and combined with the original score (SSC and FSS), which are defined in the next section (equations 3.1 to 3.5).

3.1.1 General Algorithm

The following pipeline describes the general algorithm used to find new equivalence mappings based on previous anchor alignments. The diagrams in Figures 3.1 and 3.2 illustrate this general pipeline.

Upon receiving the two alignments as input the algorithm will use the anchor alignment to search for candidate mappings. This search is parametrized by the

Structure dimension components: radius r_1 and radius r_2, as well as direction (Ancestors, Descendants, and Maximum). The algorithm will search for those mappings within a specific radius r_1 - **Find neighborhood**. Within that radius the algorithm might chose only to consider the anchor's ancestors, descendants, or both according to an input parameter. All the possible mappings between source and target neighbors inside that radius are called **candidate mappings**.

A second radius (r_2) is used to search each the candidate's neighborhood with the same direction as before (either Ancestors, Descendants, or both). This second search will consider the pairs that are contemplated in the $Alignment_{LT}$ (with lower threshold) - **similarity pairs**. This consideration means that only pairs that have alignment scores higher that the lower threshold will be considered.

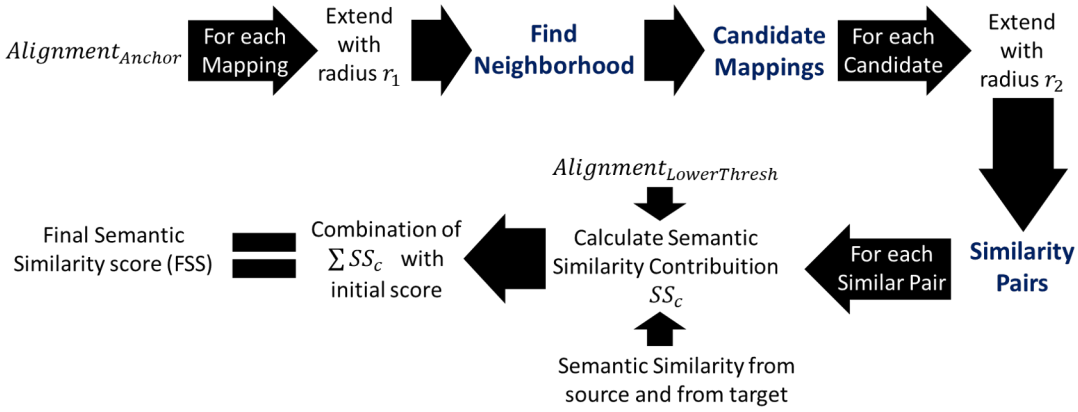
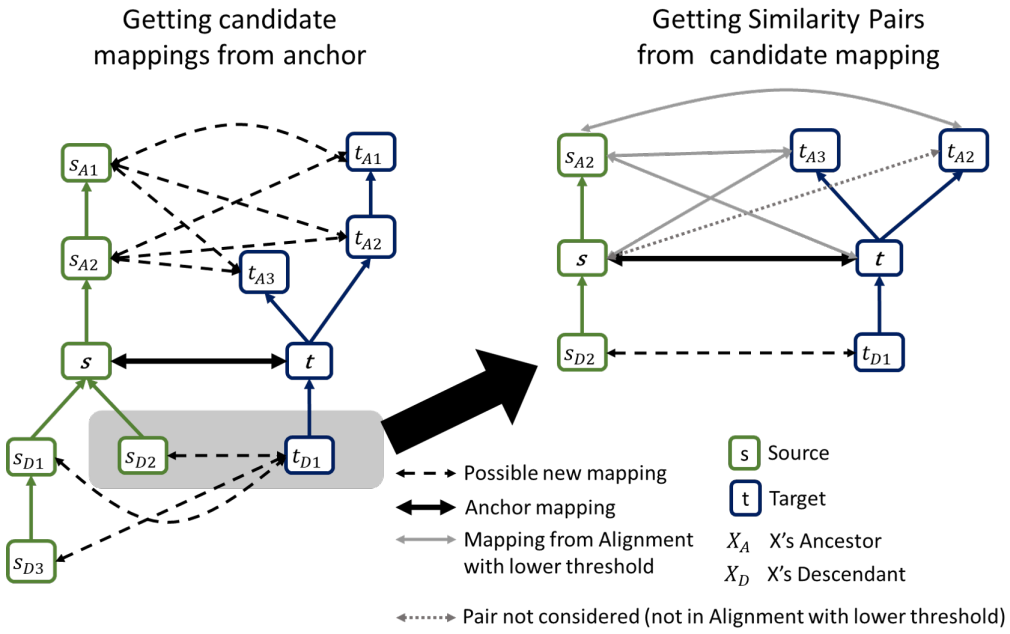


Figure 3.1: Pipeline for the General Algorithm used for the Equivalence Matchers

Each similarity pair will contribute to the new score with a semantic similarity contribution (SSc). The semantic similarity contribution takes into account the Semantic dimension parameters: IC metric and semantic similarity metric. The SSc is a combination of the semantic similarity (between the two terms from the source and the two terms from the target) with the score from the $Alignment_{LT}$. As shown in Figure 3.2 the SSc will be the combination of the semantic similarity from the source (between s_{D2} and s) with the semantic similarity from the target (between t_{D1} and t) and the score from the previous alignment - ISimL (Initial Similarity from $Alignment_{LT}$).



(a) The dashed arrows show all candidate mappings from s and t anchor. The shadowed pair will be used in 3.2(b)

(b) Similarity pairs generated from s_{d2} and t_{d1}.

Figure 3.2: General steps for the Equivalence Matchers with r_1 and r_2 equal to 2 as a schematic example.

The computation of the final score (**FSS**) for each candidate mapping (s-t) will take into account the average values of SSc as well as the score for the mapping from the $ISimL$ from s-t. FSS will be the incrementation of a score ($ISimL$) that might not be enough to pass the threshold initially with the surrounding scores weighting the semantic similarity. Therefore the candidate mappings that are approved (mappings) are the ones with enough semantic similarity to increase the lower score towards the matcher's threshold value.

As part of the methodology, there are two different ways to calculate the Semantic Similarity contribution (SSc_A or SSc_B respectively in equations 3.1

and 3.2).

$$\begin{aligned} SSc_A &= \frac{ISimL}{2} * SSsource + \frac{ISimL}{2} * SStarget \\ &= \frac{ISimL}{2} * (SSsource + SStarget) \end{aligned} \tag{3.1}$$

$$SSc_B = ISimL * Tconorm(SSsource, SStarget) \tag{3.2}$$

Both take into account the semantic similarity from the source and the target as well as the score between the similarity pairs from the $Alignment_{LT}$. After averaging the SSc the final score will be either Tconormed (equation 3.3)

$$Tconorm(a, b) = a + b - a * b \tag{3.3}$$

with the score $ISimL$ between the candidate mapping (FSS_A equation 3.4) or averaged (FSS_B equation 3.4).

$$FinalSS_A = Tconorm(\overline{WSS}, ISimL) \tag{3.4}$$

$$FinalSS_B = \frac{\overline{WSS} + ISimL}{2} \tag{3.5}$$

The general algorithm here described can be parametrized along the three aforementioned Dimensions: structure - the radii for semantic expansion, and the direction of this expansion; Semantic Similarity - the various ways to compute IC and semantic similarity; and Weighting Mechanism - the alternatives for computing SS_c and FSS. In the structure Dimension the radii r_1 and r_2 to ensure that the anchor mapping is inside the neighborhood search. The following contains the algorithms' description:

- For every anchor mapping:
  - Define its neighborhood as the set of descendants and ancestors' classes for each class involved in the mapping at a distance equal or inferior to r_1.
  - Each possible pairing between one descendant/ancestor from the source

ontology and one descendant/ancestor from the target ontology is classified as a candidate mapping.

- For every candidate mapping:
  - Define its neighborhood as the set of descendants and ancestors' classes for each class involved in the mapping at a distance equal or inferior to r_2.
  - Each mapping found in that neighborhood and also found in the alignment with lower threshold is defined as a similarity pair.
  - For each similarity pair:
    - Retrieve the mapping's score from the alignment with lower threshold
    - Compute the semantic similarity inside each ontology between the concept from the similarity pair and the concept from the candidate mapping.
    - Compute the Semantic Similarity contribution (SS_c) as a combination of the similarity mapping's score and both the semantic similarity (from each ontology), using the equations 3.1 or 3.2.
  - Average the SS_cs found in the candidate mapping's neighborhood.
  - Compute with the equations 3.4 or 3.5, the Final Semantic Similarity score (FSS), using the average SS_c and the candidate mapping's score.
  - If the FSS is equal or superior to the matcher's threshold the candidate mapping becomes part of the preliminary alignment.

3.2 Subsumption Matchers

The subsumption matchers developed in this work also need a pre-existing alignment as input to provide mappings to be used as anchors. Those equivalence anchors will allow the new algorithms to search on their neighborhood for subsumption mappings and relate them accordingly. In this section there are two

main algorithms that differ in complexity: basic semantic subsumption matcher and extended semantic subsumption matcher.

3.2.1 Basic Semantic Subsumption Matcher

This approach (*BSSM*) generates subsumption mappings based on simple reasoning. If S is equivalent to T, then all S's ancestors are marked as T's ancestors, and all T's ancestors are marked as S's ancestors (likewise for descendants). All new mappings thus obtained are assigned a score equal to the one between S and T. This pipeline takes an equivalence alignment as input with a threshold t_1. For each mappings (S, T) from that anchor, the methods illustrated in Figure 3.3 follow these operations:

- Finds all direct neighbors (ancestors and descendants) of S and T.
- Maps each $S_{Ancestor}$ (with the same score from the anchor) as T's Ancestor. Plus each $T_{Ancestor}$ is mapped as S's Ancestor with the score of the pair (S/T).
- Maps S (with the same score from the anchor) as ancestor of $T_{Descendant}$. Plus each T is mapped as ancestor of $S_{Descendant}$ with the score of the pair (S/T).

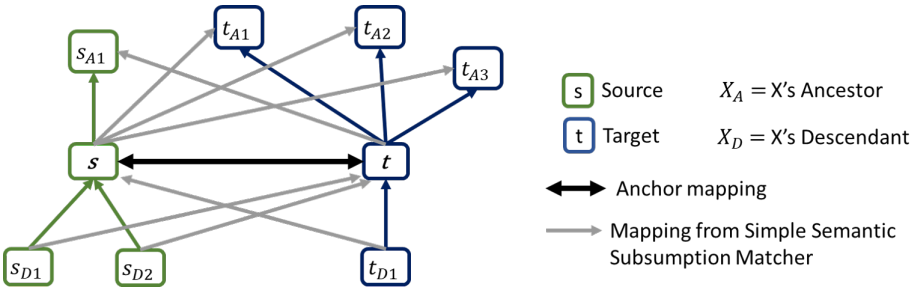


Figure 3.3: Representation of Simple Semantic Subsumption mappings. The mappings represented have the same confidence as the anchor mapping.

The output of this algorithm is equivalent to running a reasoner over the integrated ontologies. However, since integration is not always achievable or desirable, a basic matcher was implemented (Pesquita *et al.*, 2013a).

3.2.2 Extended Semantic Subsumption Matcher

The extended matcher (*ESSM*) considers the case illustrated in Figure 3.4 where it is possible to verify that the structural organization of the domain can differ between ontologies, where a sibling of S can be considered an ancestor or descendant of T due to different modeling views. A challenge then arises, the identification of the mapping's direction: should a sibling be mapped as an ancestor or descendant? Moreover, this matcher also computes new scores for the mappings, based on the similarity of the concept's labels. The algorithm runs the following steps:

1. Get the anchor mappings from the input alignment (S, T).

2. Find the direct neighborhood for both S and T plus their correspondent set of siblings. Creating a set of *candidate pairs*:

   - $S_{ancestor}$ as T's ancestor and vice-versa.
   - $S_{descendant}$ as T's descendant and vice-versa.
   - $S_{sibling}$ as either T's ancestor or descendant and vice-versa.

3. For each candidate pair, (S_i, T) and (S, Tj), generate new scores:

   - get main label and set of synonyms - *label*;
   - normalize each label;
   - compute string similarity between each label of each pair;
   - calculate the new score between (S_i, T) and (S, Tj) as the maximum score between their labels;
   - if the mapping involves a sibling, identify the direction of mapping;
   - if the score $>$ threshold, the candidate pair is added to the alignment becoming a mapping.

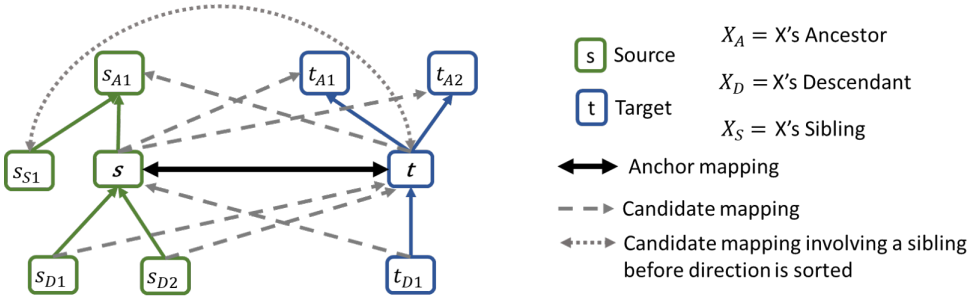


Figure 3.4: Diagram of candidate mappings identified by ESSM

Three alternatives for calculating label similarity were included in *ESSM*:

- **subsWord** (Wor) - a bag-of-words similarity method that uses all available labels (main plus synonyms). It uses a simple *Jaccard* distance (intersecting words over all words).

- **subsStemmer** (Stem) - same structure and principles of *subsWord* but the *labels* suffer a specific lemmatization using a *Stemmer*. The *Snowball Stemmer* is used to reduce each word to its morphological root, allowing metabolic and metabolism to be regarded as equivalent labels. For example allowing the mapping between *disorder of amino acid and other organic acid metabolism* and *amino acid metabolic disorder*.

- **subsString** (Str) - a string similarity method that removes stop words and non-alphanumeric characters ('and', 'by', 'has', 'is', 'non', 'or', 'of', 'to', '(', '-', '_', ')', 'as', 'with', 'type', 'on', 'in', 'at', 'part', and ','). This enables the detection of subsumption mappings such as *artery* and *part of artery* or even *hypersensitivity reaction* and *hypersensitivity reaction type IV*. In theory any string similarity metric could be used but the chosen implementation (see Chapter 4) uses ISub (Stoilos *et al.*, 2005)

Since string similarity is used to find equivalences, the words removed are the ones that would not interfere with the concept's exact meaning. When searching for subsumptions there is a need for a hierarchical search. Therefore the range of words to remove can be expanded to words that would interfere with an equivalence match but not with a subsumption match. For example *part of* in *artery*

and *part of artery* which would not be an equivalence but should be encountered for subsumption.

The example *artery* and *part of artery* can also be used for the choice of direction of the subsumption relation when dealing with siblings. As explained before, without the involvement of siblings the direction decision is very straightforward: A's ancestors will be matched as a ancestors of B; A's descendants will be mapped as descendants of B; and vice-versa. When siblings are added to potential mappings, the assumption of hierarchical relationships is not straightforward anymore: A's siblings might be B's parents or children. This problem arises when dealing with ontologies that have different points of view because they consider different levels of detail for the same domain.

Upon finding the best similarity, the most specific entity will be the longest one. This means for the pair of labels with the best similarity, the entity considered the ancestor is the one with less words and/or smaller length. This assumption is supported by the fact that many biomedical ontologies use regular lexical structures in their labels that can be explored (Quesada-Martínez *et al.*, 2013).

Chapter 4

Evaluation Methodology

This section describes the methodology followed to implement and evaluate the matching algorithms presented in Chapter 3. The first part will cover the Data Resources employed, followed by the actual evaluation pipelines for both equivalence and subsumption, in that order.

4.1 Data Resources

The Ontology Alignment Evaluation Initiative (OAEI) is an international initiative that evaluates ontology matching systems (Achichi *et al.*, 2016). The main goal of OAEI is the evaluation of ontology matching systems, including comparing their results with reference alignments produced for this effect. Additionally, OAEI assesses strengths and weaknesses of the systems and promotes communication among algorithm developers, thereby contributing to the improvement of the current ontology alignment panorama. OAEI was created in 2004 as an yearly international event organized to perform these evaluations. This initiative provides ontologies, or fragments thereof, to be matched, as well as reference alignments that might contain curated mappings. All data resources used in this dissertation came from OAEI 2016.

In order to test the *Equivalence Matchers*, four ontologies were used:

- Foundation Model of Anatomy ontology (*FMA*) focuses on the representation of the phenotypic structure of the human body - anatomy (Rosse &

Mejino Jr, 2008);

- National Cancer Institute Thesaurus (*NCI*) implements reference terminology for the National Cancer Institute by covering a vast domain of vocabulary, from clinical care, public information, and administrative activities, to translational and basic research (Golbeck *et al.*, 2011);

- Systematized Nomenclature of Medicine, Clinical Terms (*SNOMED CT*) is a comprehensive and precise clinical health terminology developed to accommodate the diverse needs and expectations of the worldwide medical profession (Donnelly, 2006);

- Adult Mouse Anatomy (*MA*), which is an anatomical ontology for adult mouse terms (Hayamizu *et al.*, 2005).

Those ontologies are used either as a whole or as parts which were made available by OAEI. The available whole ontologies are NCI and FMA, there are also smaller fragments for both ontologies plus a small and a large fragments from SNOMED. The smallest fragments are the Mouse and Human which are fragments that only consider the anatomy portions of MA and NCI respectively.

The evaluation of the *Subsumption Matchers*' was made through the usage of:

- Human Phenotype Ontology (*HP*), which is a description of phenotypic abnormalities, by providing a standardized vocabulary of the ones encountered in human diseases (Köhler *et al.*, 2013);

- Mammalian Phenotype Ontology (*MP*) is the representation of the characteristics of mammalian organisms manifested through either lifespan and/or development - observable morphological, physiological, and behavioral (Smith & Eppig, 2012);

- Human Disease Ontology *DOID* is a comprehensive controlled vocabulary for human diseases (Schriml *et al.*, 2012);

- Orphanet and Rare Diseases Ontology *ORDO*, an ontology dedicated to rare diseases available in multiple languages (Vasant *et al.*, 2014).

The evaluation of the matching for those ontologies is done through comparison with reference alignments that were also available in OAEI. The reference alignment for the Mouse-Human pair is the most curated alignment. The anatomy track which is the matching of Mouse to Human (MH) has been done since 2005 and the reference alignment has been made public in 2008. FMA, NCI, and SNOMED are matched for every combination of their small fragments, and combination of the large fragments. Those reference alignments are generated from the Unified Medical Language System (UMLS) Metathesaurus. UMLS is the most comprehensive effort for integrating independently-developed medical thesauri (Jimenez-Ruiz *et al.*, 2010).

With regard to the subsumption track, the reference alignments available for HP-MP (HM) and DOID-ORDO (DO) are the set of extracted mappings automatically generated by BioPortal. These reference alignments are not of high quality since they are automatically generated without any manually curation. Even OAEI prefers to use silver standards (mapping found by at least two/three systems) instead of those reference alignments in their evaluation (Achichi *et al.*, 2016).

OAEI includes different tracks to evaluate different aspects of ontology matching. The availability the Mouse-Human reference alignment has made it possible for all the systems to find better solutions for this particular ontology matching problem. This pair is particularly different from all the others for its size, specificity, and the curated reference alignment.

As stated, the ontologies used can be fragments of the actual ontology. In Table 4.1 the ontologies used for testing the equivalence matcher are presented. When creating a fragment of an ontology to be matched to another there is a need to find sections from both ontologies that overlap. This need is important to guarantee that there are mappings to be found. The lack of this procedure can lead to the absence of mappings.

Table 4.1 presents the general aspects of the ontologies used in the Equivalence Matchers. The pair Mouse-Human has a different profile than the others. Besides being the smallest pair, MH also is the more vertical one. Human has a Maximum Depth of 11 and Mouse of 6, for their size the depth should be smaller when compared to the other pairs.

Table 4.1: Characteristics of the ontologies with for the equivalence matchers

Ontology from OAEI	Number of Classes			nº of Roots	Maximum Depth
	Total	With a Single Child	With >25 Children		
Human	3304	209	28	8	11
Mouse	2743	8	0	3	6
FMA small overlapping NCI	3696	895	13	4	18
FMA small overlapping SNOMED	10157	1139	70	4	18
FMA whole ontology	78988	118	243	4	20
NCI small overlapping FMA	6488	856	28	16	12
NCI small overlapping SNOMED	23958	2290	115	19	14
NCI whole ontology	66724	4744	372	19	14
SNOMED small overlapping FMA	13412	2404	34	10	25
SNOMED small overlapping NCI	51128	11873	375	17	23
SNOMED extended overlapping FMA and NCI	122464	22892	917	19	26

4.1.1 AgreementMakerLight - AML

The algorithms developed for both types of matching were implemented as an extension of AML (Faria *et al.*, 2013a), but their theoretical foundations are independent from this platform, since they could be implemented within other extensible ontology matching systems as well or independently. AML was chosen because it is a top performing system in the area of Ontology Matching in biomedical ontologies and easily extensible. The Semantic Similarity Library (SML) was employed to support semantic similarity calculations. This chapter also presents the steps taken to integrate the algorithms with AML and SML.

AML focuses on computational efficiency handling very large ontologies. After loading the ontologies, it uses a variety of matchers and filters depending on the characteristics of the ontologies loaded, e.g. size (Faria *et al.*, 2013a). This system follows the pipeline illustrated on Figure 4.1 .

AML uses the OWL API not only to load the ontologies but also to parse them into its data structures. In particular it contains a data structure called the Lexicon that stores lexical information, and the RelationshipMap that stores structural information. As presented on Figure 4.1, there are many matchers implemented in AML. For the sake of readability only the matchers relevant to this dissertation will be presented:

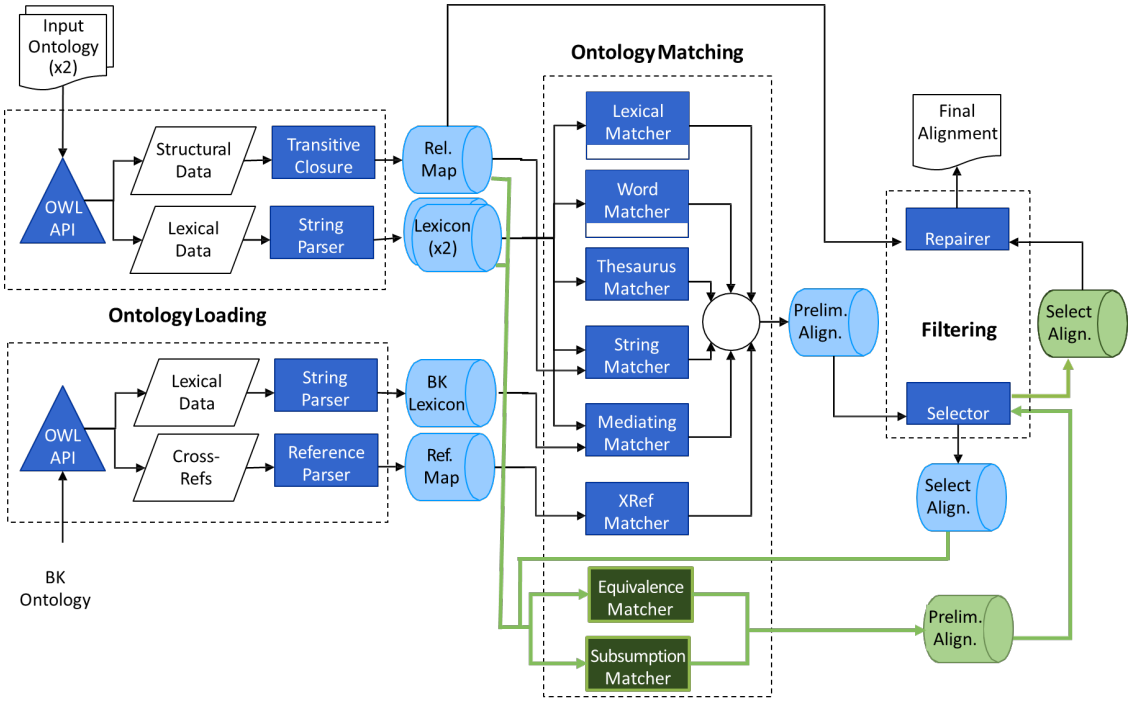


Figure 4.1: Integration of developed Matchers into AML's pipeline*

* There are four parts to this pipeline. The three parts from AML: ontology loading - input or background knowledge (BK) ontologies are parsed and loaded into AML's data structures; ontology matching - matchers generate candidate mappings latter combined into a preliminary alignment; and filtering - problem-causing mappings (e.g. cardinality) are excluded from the previous alignment to produce a final alignment. While the fourth part is the implementation of the new matchers: this implementation occurs after the first two steps and the selector part of the third step; after running these matchers there is the filtering step is done in fullness. In green are the matchers created plus the extra steps included for the evaluation.

- **LexicalMatcher** - searches for literal full-name matches between the Lexicons from both ontologies;
- **WordMatcher** - finds matches by computing overlapping words and scores the matchers by amount of overlap;
- **StringMatcher** -explores string similarity by computing the similarity between the Lexicon entries (of the pair to match).

Filtering is responsible for dealing with cardinality problems by using one of three selectors (strict, permissive, or hybrid). The selector receives an alignment

as input and filters it based on a threshold, removing every mapping with a score under the threshold. The difference lies on how the selector handles the cardinality conflicts, upon finding two conflicting mappings from a concept A to both B and C, with different or equal scores:

- **Strict** - keeps only the mapping with the higher confidence score, in case of the same score, the selector keeps only one, chosen randomly;
- **Permissive** - keeps all the mappings that have the highest confidence score;
- **Hybrid** - between the selector's threshold and 0.75 behaves as the permissive selector, and allows conflicting mappings over 0.75 by keeping every match.

4.1.2 Semantic Similarity Implementation

The Semantic Measures Library (SML) is an open source Java library devoted to the computation of semantic measures (Harispe *et al.*, 2015). SML can be used to compute semantic similarity, semantic relatedness, or semantic distance, between others. The implementation of SML depends on its own ontology loading since just like AML it uses special data structures:

- URIFactory is the main data structure that will store all the concepts in a URI form.
- SM_Engine is a data structure with the method to compute the semantic similarity between two terms with the measure and IC intended.

The parsing is made through an external parser which does not allow large ontologies. That problem lead to the necessity to implement algorithms that would compute the semantic similarity. The implementation of the algorithms for the computation of semantic similarity ($mySSM$) for larger ontologies was done using AML and its own data structures. $mySSM$ implementation does not require a second loading and parsing of the ontologies reducing the time needed for the semantic similarity computation.

4.2 Equivalence Matcher's Evaluation Pipeline

The choice of AML was made not only by the ease and availability of this software but also for the fact that this is and has been a top performing participant in OAEI for the past four years. AML has evolved with new algorithms and strategies every year, becoming a powerful solution for ontology matching. Particularly in biomedical ontologies matching, AML explores external ontologies and resources as background knowledge, increasing its performance considerably. However, these external resources are not always available or simple to use. Thus our evaluation approach takes as baseline a simpler, generic matcher, applicable in any biomedical ontology matching scenario and focused on lexical similarity. This is not a poor-performing approach and in fact outperforms several OAEI participants.

Tables 4.2 and 4.3 summarize the way the results are structured. The ontology pairs are divided by size: MH is small; FNs and FSs are medium; SNs is large; while FNw, FSw, and SNw are very large. The size nomenclature is not always coherent in literature therefore the one used is from AML's size evaluation. To establish the best strategies the *Baseline* was used on the small and medium sized pairs: Mouse-Human, FMA-NCI small fragments, and FMA-SNOMED small fragments. This pipeline consists in only LM extended with SM, and finally a Selector. The strategies were then grouped into the previous dimensions. Each parameter in each dimension is evaluated by grouping the results obtained with all strategies that use that parameter, selecting only the strategies with highest F-measure (the top 20) and averaging the performance measures. These averages are then compared to their correspondent baseline alignment (LM + SM + Selector). Those comparisons would lead to the selection of the best strategies that would be evaluated with AML's complete pipeline for the same pairs plus NCI-SNOMED small fragments (NSs), FMA-NCI large fragments (FNw), FMA-SNOMED large fragments (FSw), and NCI-SNOMED large fragments (NSw).

Table 4.2: Ontology pairs presented in each table designed for each dimension. *

Pipeline	Evaluation type	Ontology Pairs							Tables
		MH	FNs	FSs	SNs	FNw	FSw	SNw	
Baseline	Average top 20 F-measures**	+	+	+	-	-	-	-	4.3
	Top strategies	+	+	+	-	-	-	-	5.4
Complete	Top strategies	+	+	+	+	+	+	+	5.5

* The pipeline used was the *Baseline* (LM + SM + Selector). MH: Mouse-Human, FN: FMA-NIC, FS: FMA-SNOMED, SN: SNOMED-NCI; s: small section of the ontology pair, w: whole or large section of the pair.

** The average of the strategies that hold the best 20 F-measure results for each one of the previously presented Dimensions.

Table 4.3: All the strategies and their respective results tables.

Pipeline	Structure		Semantic Similarity		Weighting Mechanism		
	Direction	Distance	IC	Measure	SSc	FSS	threshold
Baseline	Ancestors Descendants Maximum	1 3 5	Seco Zhou Sanchez	Resnik J. Conrath Lin simGic	SSc_A SSc_B	FSS_A FSS_B	0.6 0.7 0.8
Table*	5.1		5.2		5.3		

* Each table show the average of the best 20 F-measure results filtered by Dimension.

4.3 Subsumption Matcher's Evaluation Pipeline

The evaluation for the *Subsumption Matchers* was based on the sets of reference alignments for the pairs aligned; plus an extra manual evaluation. Only the HP-MP and DOID-ORDO were evaluated since those pairs were the ones used for the subsumption track in OAEI. These matchers use three dimensions: the semantic similarity, the string similarity and the selection used. Each strategy pipeline has: (i) the unmodified result of running AML's main algorithm as anchor, with (ii) one of the four ($BSSM$ and the three similarity options for $ESSM$) subsumption matchers described before (using a threshold tMatcher), and (iii) one of the three selectors described earlier (using a threshold tSelector and an additional threshold tHybrid for the Hybrid selector). Both tMatcher and tSelector take the values 0.6, 0.7, or 0.8 while the tHybrid will take the values 0.7, 0.75, or 0.8.

The manual evaluations were performed by: first checking on BioPortal (Whetzel *et al.*, 2011) for their synonyms and annotations contained in the ontologies in

question, further consulting one or two medical dictionaries[1,2], and finally search for scientific papers that would somehow contextualize the previous information (Wheeler *et al.*, 2007). If after all the previous steps there was still doubt the pairs were marked as not subsumption.

In terms of manual evaluation, 30 random mappings were selected from the resulting alignment for each strategy - in an attempt to create a similar evaluation to OAEI. Each pair was given an assessment regarding the direction of the subsumption. After the elimination of duplicates, the final lists contained 151 pairs (DOID and ORDO) and 181 pairs (HP and MP). The new assessment takes into account the direction of the relationships. A pair could be positive for *subsumption* (**is_a** or **part_of** relationships), or negative (cases of *equivalence*, *incorrect*, or *Different Subsumption*). The *Different Subsumption* evaluation means that the matcher selected a subsumption mapping but incorrectly chose the direction of the subsumption. To compare the behavior of the different strategies, the assessment would fall back to the precision of subsumption mappings:

$$\frac{\text{subsumption}}{\text{subsumption} + \text{equivalence} + \text{incorrect} + \text{Different Subsumption}} \tag{4.1}$$

[1]http://medical-dictionary.thefreedictionary.com
[2]http://www.online-medical-dictionary.org

Chapter 5

Results and Discussion

The results and their discussion are organized as follows: first regarding the Equivalence Matchers, followed by the results and discussion from Subsumption Matchers.

5.1 Equivalence Matchers

Each Dimension (Structure, Semantic Similarity, and Weighting Combination) was evaluated independently. For each Dimension, the top 20 F-measure were averaged to allow a general overview of the impact of each Dimension parameters in performance of the algorithm. An overall evaluation of the combination of the best parameterizations of each Dimension was also performed. In this overall evaluation, the results using *Baseline* pipeline as input alignments are presented for the small and medium pairs (MH, FNs, and FSs) and the AML full pipeline are used as anchor for all the ontology pairs.

This last part enables the comparison of the strategies as well as the impact of the anchor alignment. Before the Equivalence results are presented it is important to highlight that given AML's optimized performance and the fact that the *Baseline* itself outperforms several OAEI participants, seemingly small improvements can be considered relevant.

5.1.1 Structure

The parameters grouped in the Structure dimension are direction of the semantic expansion and the radius of the expansion. Table 5.1 describes the Structure dimension results using the *Baseline* as input alignments. For the medium size pairs (FNs and FSs), the Ancestors with radius 3 is present the best results for F-measure, followed by radius 5. For the MH the radius 3 is still in the top but now the best direction is Maximum, followed by Descendants with radius 3 as well. For FNs and MH all the strategies show an improvement in terms of F-measure regarding the baseline. On the other hand, FSs only presents better scores when dealing with Ancestors.

Theoretically, the best performance should be the Maximum which means that the correct scores are the ones with higher scores. The fact that FSs and FNs shows higher scores for Ancestors means that wrongly identified mappings from Descendants (that are in conflict) will have higher scores than Ancestors. Plus, those fragments are specifically for anatomy, meaning that not only the string similarity for correct mappings will be more accurate but the computation of semantic similarity as well.

Table 5.1: Average Precision, Recall, and F-measure from SML for the best 20 aggregated by Structure, which includes direction and distance*.

Ontology Pairs	Structure		Average best 20 (%)		
	Direction	Radius	Precision	Recall	F-Measure
FNs	Baseline		96.686	84.736	90.317
	A	1	96.669	84.783	90.336
	A	**3**	**96.611**	**84.896**	**90.375**
	A	5	96.662	84.840	90.366
	D	1	96.687	84.744	90.322
	D	3	96.683	84.765	90.333
	D	5	96.683	84.765	90.333
	M	1	96.687	84.744	90.322
	M	3	96.684	84.787	90.345
	M	5	96.684	84.777	90.340
FSs	Baseline		94.393	67.325	78.594
	A	1	94.213	67.463	78.624
	A	**3**	**93.624**	**67.907**	**78.717**
	A	5	93.871	67.727	78.683
	D	1	94.393	67.325	78.594
	D	3	94.393	67.325	78.594
	D	5	93.819	67.485	78.501
	M	1	92.762	67.797	78.334
	M	3	93.037	67.883	78.491
	M	5	89.994	68.322	77.667
MH	Baseline		97.906	77.111	86.273
	A	1	97.863	77.399	86.436
	A	3	97.659	77.562	86.457
	A	5	97.669	77.451	86.392
	D	1	97.514	77.684	86.476
	D	3	97.461	77.847	86.555
	D	5	97.461	77.847	86.555
	M	1	97.655	77.611	86.485
	M	**3**	**97.487**	**77.878**	**86.585**
	M	5	97.459	77.829	86.544

* FNs: FMA-NCI small fragment, FSs: FMA-SNOMED small fragment, MH: Mouse-Human; A: Ancestor, D: Descendant, M: Maximum.

5.1.2 Semantic Similarity

Table 5.2 show the results from the Dimension of Semantic Similarity. For the medium size pairs, Zhou (Z) combined with Resnik(R) produced the best average results, while for MH that happens for Seco (S) with Resnik.

The decision of best measure is consensual, Resnik is the measure that holds the best scores regardless of the ontology pair. By looking at the results of the medium pairs, the measure used has more impact than the IC approach, the difference between Zhou/Resnik and Seco/Resnik are 0.003% for FNs and 0.004% for FSs, while for MH the difference is 0.144% between the same strategies. This information points to a decision towards the Seco/Resnik strategy, plus this strategy shows an improvement towards the baseline of 0.044% for FNs, 0.056% for FSs and 0.296% for MH.

Table 5.2: Average results for the best 20 strategies aggregated by Semantic Similarity.*

Ontology Pair	Semantic Similarity		Average best 20 (%)		
	IC	SSMea	Precision	Recall	F-Measure
FNs	Baseline		96.686	84.736	90.317
	S	JC	96.687	84.757	90.330
	S	L	96.650	84.787	90.330
	S	R	96.635	84.847	90.358
	S	sG	96.669	84.792	90.342
	Sa	JC	96.664	84.777	90.331
	Sa	L	96.684	84.773	90.337
	Sa	R	91.453	84.767	87.970
	Sa	sG	96.671	84.783	90.337
	Z	JC	96.681	84.769	90.334
	Z	L	96.683	84.769	90.335
	Z	**R**	**96.660**	**84.834**	**90.361**
	Z	sG	96.673	84.779	90.336
FSs	Baseline		94.393	67.325	78.594
	S	JC	94.011	67.530	78.598
	S	L	94.041	67.510	78.595
	S	R	94.241	67.481	78.646
	S	sG	94.214	67.466	78.627
	Sa	JC	94.066	67.521	78.611
	Sa	L	93.454	67.682	78.502
	Sa	R	90.923	67.451	77.407
	Sa	sG	94.170	67.511	78.642
	Z	JC	93.437	67.719	78.520
	Z	L	93.589	67.695	78.559
	Z	**R**	**94.231**	**67.493**	**78.650**
	Z	sG	94.142	67.506	78.628
MH	Baseline		97.906	77.111	86.273
	S	JC	97.602	77.590	86.451
	S	L	97.669	77.621	86.498
	S	**R**	**97.610**	**77.774**	**86.569**
	S	sG	97.501	77.795	86.539
	Sa	JC	97.430	77.857	86.548
	Sa	L	97.692	77.555	86.466
	Sa	R	89.948	78.243	83.661
	Sa	sG	97.426	77.777	86.498
	Z	JC	97.677	77.493	86.422
	Z	L	97.703	77.517	86.447
	Z	R	97.648	77.517	86.425
	Z	sG	97.541	77.691	86.490

* FNs: FMA-NCI small fragment, FSs: FMA-SNOMED small fragment, MH: Mouse-Human; IC: Information Content, S: Seco, Sa: Sanchez, Z: Zhou; SSMea: Semantic Similarity Measure, S: Seco, Z: Zhou, Sa: Sanchez; JC: Jiang Conrath.

5.1.3 Weighting Mechanism

Table 5.3 shows the last Dimension evaluated. Again, in MH the best parameters are different than in the other ontology pairs. For the medium pairs, FSS_B show the best results for F-measure when paired with SSc_A (90.392% for FNs and 78.737% for FSs) and second best when paired with SSc_B (90.364% for FNs and 78.693% for FSs). It is important to note that regardless of the pair, the approaches that involve FSS_B always score better than the F-measure's baseline.

Table 5.3: Average Precision, Recall, and F-measure for the best 20 aggregated by Weighting Mechanism*.

Ontology Pair	Weighting Mechanism		Average best 20 (%)		
	SSc	FSS	Precision	Recall	F-Measure
FNs	Baseline		96,686	84,736	90,317
	A	A	96,164	84,890	90,176
	A	**B**	**96,679**	**84,873**	**90,392**
	B	A	93,252	84,863	88,858
	B	B	96,635	84,857	90,364
FSs	Baseline		94,393	67,325	78,594
	A	A	90,484	68,205	77,775
	A	**B**	**93,901**	**67,791**	**78,737**
	B	A	87,273	68,047	76,410
	B	B	93,808	67,776	78,693
MH	Baseline		97,906	77,111	86,273
	A	**A**	**97,074**	**78,236**	**86,641**
	A	B	97,669	77,798	86,608
	B	A	96,692	77,729	86,178
	B	B	97,698	77,764	86,598

* FNs: FMA-NCI small fragment, FSs: FMA-SNOMED small fragment, MH: Mouse-Human.

There is one last parameter that has not been yet accounted for, the threshold used for the matchers. This parameter is not a part of the algorithm and ontology matching systems typically set this as a manual input. To elucidate if a given threshold was more appropriate, the percentage of each tested threshold (0.6, 0.7, and 0.8) in the top 20 approaches was calculated. Figure 5.1 shows that, in the top 20 F-measure results there is a major presence of the threshold 0.7 in comparison to the other thresholds and regardless of the ontology pair.

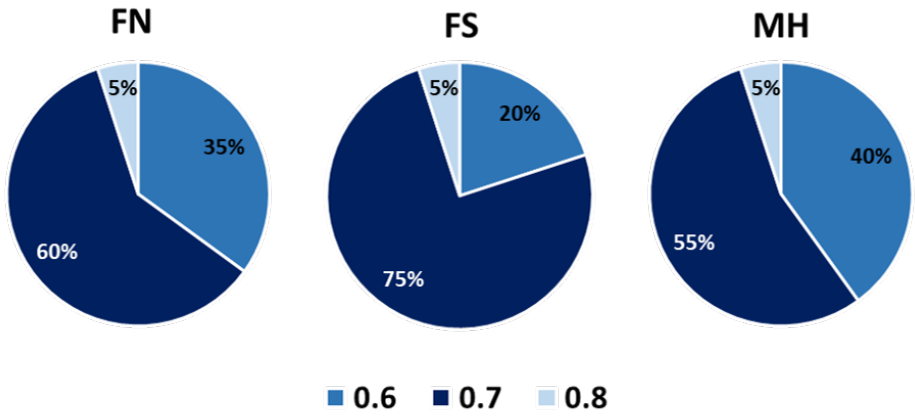


Figure 5.1: Percentage of thresholds in the top 20 of the best F-measure results filtered by ontology pair

5.1.4 Overall Results

Until now all the results presented were through averages of a selected top of strategies rather than specific strategies. The next two tables present the strategies that combine the best parameters found by the Dimension results analysis. Table 5.4 presents the best strategies using the Baseline alignment and Table 5.5 presents the best strategies using AML's complete pipeline.

Table 5.4: Results from the top strategies with the Baseline pipeline medium and small pairs of ontologies*.

OP	Structure	SS	Weighting Mechanism	P	R	F
FNs	Baseline			96.686	84.736	90.317
	Ancestors/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	96.688	84.773	90.339
	Ancestors/3	Seco/Resnik	$SSc_B/FSS_B/0.7$	96.694	84.922	90.426
FSs	Baseline			94.393	67.325	78.594
	Ancestors/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	94.373	67.358	78.609
	Ancestors/3	Seco/Resnik	$SSc_B/FSS_B/0.7$	93.836	67.956	78.826
MH	Baseline			97.906	77.111	86.273
	Ancestors/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	97.906	77.111	86.273
	Ancestors/3	Seco/Resnik	$SSc_B/FSS_B/0.7$	97.908	77.177	86.315

* OP: Ontology Pair, FNs: FMA-NCI small fragment, FSs: FMA-SNOMED small fragment, MH: Mouse-Human; SS: Semantic Similarity, tresh: threshold, P: Precision, R: Recall, F: F-measure.

It is important to note that in the Dimension analysis the medium pairs always lead to the same parameters, which can not be said for MH. The latter has shown some discrepancies when compared to the medium pairs which was expected since its size and granularity differ from the medium pairs. Moreover, MH is the longest running biomedical track in OAEI and has a fully manual and high quality reference alignment. This information is important to account for the fact that AML has evolved through the years in order to perfect its matching techniques. This pair represents a very specific sub-domain (anatomy) where most of the labels for correct mappings should be mapped through Lexical and String similarity. Plus the availability of a curated reference alignment that has been public for years has helped AML's ability to find mappings through those techniques.

Table 5.5: Results from the top strategies with the complete pipeline for all the ontology pairs*

OP	Structure	Semantic Similarity	Weighting Mechanism	Precision	Recall	F-Measure
FNs	Baseline			95,842	90,953	93,333
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	95,842	90,953	93,333
FSs	Baseline			92,283	76,203	83,476
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	92,264	76,203	83,468
MH	Baseline			95,044	93,602	94,317
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	95,044	93,602	94,317
SNs	Baseline			91,378	73,649	81,561
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	91,359	73,655	81,557
FNw	Baseline			80,531	88,086	84,139
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	80,476	88,086	84,109
FSw	Baseline			68,519	71,009	69,742
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	66,025	71,142	68,488
SNw	Baseline			86,161	68,698	76,445
	A/3	Seco/Resnik	$SSc_A/FSS_B/0.7$	86,074	68,704	76,415

* OP: Ontology Pair, FNs: FMA-NCI small fragment, FSs: FMA-SNOMED small fragment, MH: Mouse-Human; SNs: SNOMED-NCI small fragment; FNw: FMA-NCI large fragment, FSw: FMA-SNOMED large fragment, SNw: SNOMED-NCI large fragment.

Recapitulating, the best strategies from the previous results are Ancestors combined with radius 3 for Structure, Seco with Resnik, for Semantic Similarity, and SSc_A/FSS_B for Weighting Mechanism, with the 0.7 threshold. Considering

the fact that for Weighting Mechanism there is a smaller difference in results when comparing the different approaches, table 5.4 shows the results with these parameters as well.

When looking at the results with the Baseline anchor, all the strategies show better results than the baseline (except MH with SSc_A). The improvement ranges from 0.042% (MH), to 0.109% (FN) and 0.232% (FS). All the best results come from the Weighting Mechanism of $SSc_B/FSSB$.

Since a combination of the best independently analyzed parameters may not correspond to the best overall strategy, the best eight overall strategies for each small/medium ontology pairs using the *Baseline* were also applied with AML's full pipeline. Figure 5.2 presents the difference between the best strategy's F-measure and their correspondent baselines for each ontology pair. Each ontology pair has two values corresponding to the same strategy either with the Baseline or the complete AML pipeline.

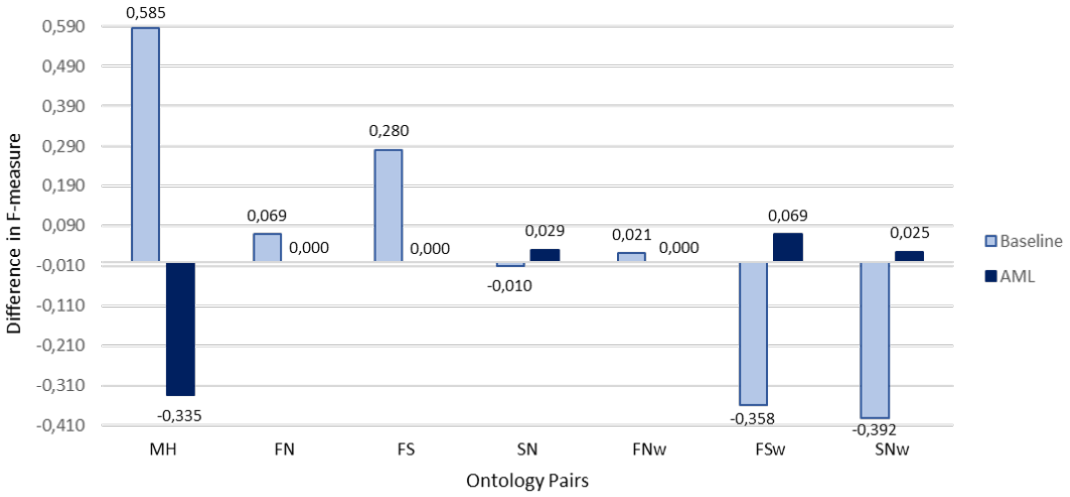


Figure 5.2: Comparison of the best strategies using the baseline and AML's full pipeline*

* Strategies used: MH - Maximum, radius 1, Seco/Jiang Conrath, SSc_A/FSS_B, and threshold of 0.6; FN, FS and SN's - Ancestors, radius 3, Seco/Resnik, SSc_B/FSS_A, and threshold of 0.7; FNw, FSw and SNw - Ancestors, radius 3, Seco/Jiang Conrath, SSc_A/FSS_A, and threshold of 0.8.

For the small and medium pairs there is a major improvement of F-measure

when dealing with the Baseline as the anchor alignment. Although, when the complete AML pipeline is used MH show a major decrease of 0.33% of F-measure and there are no new mappings for FN. For FS there is a slight improvement of 0.07% in F-measure.

As for the larger pairs, there is a clear decrease in F-measure when dealing with the Baseline (except FNw) but a small increase when the complete pipeline is used. This is a clear indication of the impact of the quality of the input alignments. The *Baseline* alignment, by virtue of being based on lexical similarity has lower recall and is prone to identify incorrect mappings that have similar labels. When these erroneous mappings are used as anchors the strategy suffers.

To elucidate the performance of the proposed approach regarding the limitation of the search space, meaning the radius around the anchors, the recall restricted to these areas was calculated. Figure 5.3 shows that the strategies found at least 66.7% of the mappings in the reference alignment within the search space. The results are quite satisfying for this dissertation scope since when recall is 0% it means that in that search space there were actually no correct mappings to be found. This figure is illustrative of the importance on the anchor alignment that will define the search dimension and the efficacy of the strategies that score between 66.7% and 100%.

Is important to understand the AML suffered many changes from 2016 to 2017's OAEI edition and the version used in this work was a beta version of the one submitted in 2017 but different from 2016. For the sake of reasoning, every time AML was used as input for SSM that same version was used including when AML was presented as baseline in this section. The disparities in F-measure results (Figure 5.4) between AML's participation in 2016, 2017, and the ones used as baseline through this chapter are due to the previous reason. Both AML and LogMap are top performing systems with high F-measure results meaning that the increase of 0.1% would be a relevant improvement in one of those systems. Decreases in F-measure are not uncommon since the implementation of strategies that benefit one pair or the system's running capability can lead to the loss in F-measure for another pair (e.g., AML increased FSs's F-measure by 7.3% by loosing 0.1% in FNs).

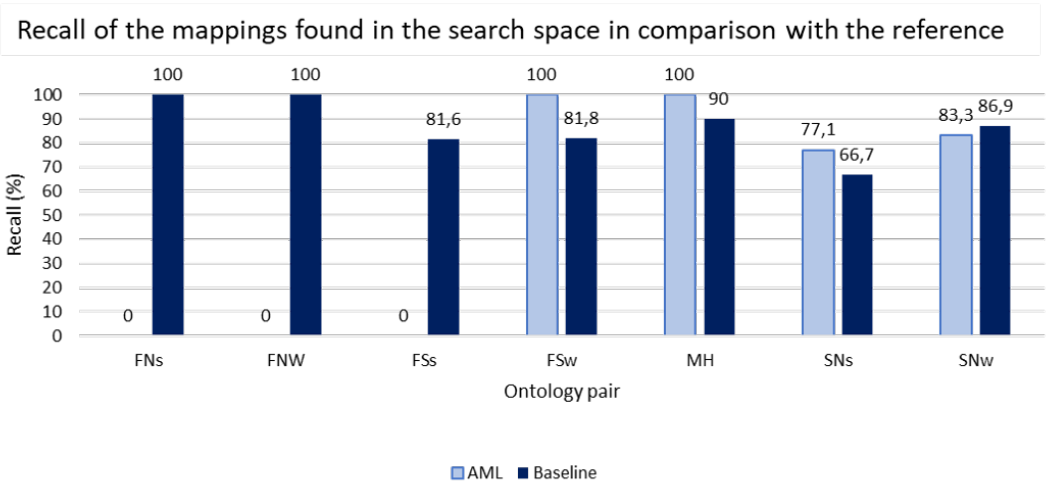


Figure 5.3: Recall in the search space. This Recall is the percentage of mappings found by the strategies (from Figure 5.2) in comparison to the ones in that scope that could actually be found (existing in reference) *

* The 0% Recall values coincide with scopes whereas the number of mappings in the reference was 0.

To propose an overall comparison of the proposed approach and the state of the art, the obtained results were compared to the OAEI 2016 and 2017 edition results for AML, LogMap, and LogMapLt. LogMap and LogMapLt are systems with similar complexity as AML and Baseline, respectively. Figure 5.4 has two goals. First, to show the difference in improvement during one year, and secondly to present comparable pipelines to consolidate the results presented. In terms of evolution, the only big difference is from LogMapLt for MH where F-measure increases 10%. As stated before, the MH pair is special for its size and reference alignment availability, plus this increase in 2017 implied a loss of 3.4% for the LogMap system.

Second, Figure 5.4 aims to compare the results from this dissertation with LogMap. Recapitulating, the Baseline input alignment does not include any Background Knowledge which was a breakthrough for the ontology matching systems by incrementing the lexical information available. Just like Baseline's pipeline, LogMapLt also lacks BK thus the similarity in terms of complexity introduced before. Even though the implementation of Semantic Similarity Match-

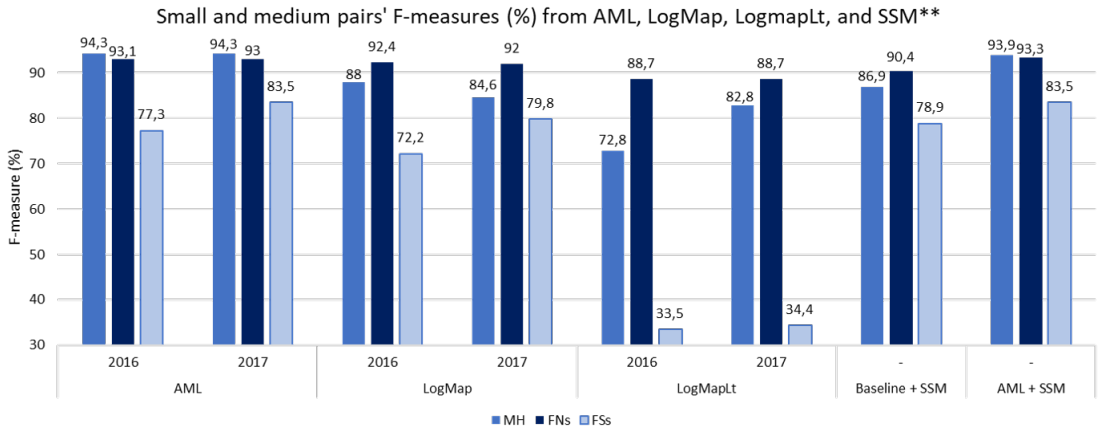


Figure 5.4: Comparison of F-measures (%) results from AML, LogMap, and LogMapLt in OAEI from 2016 and 2017 as well as the results from SSM* using either Baseline or AML as input.

* The strategies used were the same as in Figure 5.2. The results shown are from the small and medium sized pairs.

** SSM: Semantic Similarity Matcher; MH: Mouse-Human, FN: FMA-NCI, FS: FMA-SNOMED; s: small fragments.

ers does not directly improve AML's complete pipeline results, these generalist strategies have shown to be valuable for the Baseline. When comparing the results from AML with SSM with results with LogMap there is improvement in the F-measure of all ontology pairs. On another hand, Baseline with SSM always score better than LogMapLt and LogMap in 2016 for the FSs Pair. The fact that the Baseline + SSM more than doubles the score from LogMapLt for FSs pair is a very good indicator of the applicability and generality of these matchers.

5.2 Subsumption Matchers

For the subsumption matchers, the evaluation focused on comparing the matcher strategy (BSSM, Wor, Stem, and Str), and their combination with different selection approaches. Both the matcher and the selector have thresholds that will be parameters for the tests, and the case of other threshold is added, that will substitute the fixed 0.75 threshold.

Each strategy is identified by the type of matcher plus the threshold used for

that matcher (e.g., Sem0.6), and the selector plus threshold (e.g., Per0.6). In the case of the Hybrid selector, it will be just like the others with the addition of _threshold for the Hybrid threshold (e.g., Hyb0.6_0.75).

Table 5.6 shows the number of New Mappings found in comparison to the AML baseline for each strategy, the coverage is the percentage of the ontology that is mapped (either as equivalence or subsumption). This table shows that the average coverage of ORDO is consistently smaller than the coverage of DOID

Table 5.6: New mappings and coverage for each strategy for DOID-ORDO and HP-MP tasks.

Strategy		New Mappings DO	Coverage (%) DOID	Coverage (%) ORDO	New Mappings HM	Coverage (%) HP	Coverage (%) MP
AML (baseline)		-	20.2	15.3	-	14.2	14.2
BSSM0.6	Per0.6	4465	33.4	39.3	5226	38.0	32.5
Wor0.6	Hyb0.6_0.75	73	20.7	15.6	208	14.8	15.2
	Per0.6	73	20.7	15.6	208	14.8	15.2
Stem0.6	Hyb0.6_0.75	98	20.9	15.6	393	15.5	16.3
	Per0.6	98	20.9	15.6	393	15.5	16.3
Str0.6	Hyb0.6_0.75	409	23.4	16.6	929	18.8	17.7
	Hyb0.6_0.7	434	23.5	16.6	1002	18.9	17.9

* BSSM: Basic Semantic Subsumption Matcher, Wor: SubsWord, Stem: SubsStemmer, Str: SubsString; Per: Permissive Selector, Hyb: Hybrid Selector; DO: DOID-ORDO, HM: HP-MP;

Table 5.7 shows the results discriminated by type of subsumption. 30 random new mappings were selected from the alignments produced by each strategy to be manually evaluated. After eliminating duplicates they were evaluated and the results in Table 5.7 show the precision of those new mappings. In DO, the extended semantic subsumption matching strategies all outperformed the basic semantic approach by 7 to 20% increase in precision. In HM, the *Str* approaches achieved a lower precision than the *BSSM* approach, and the other approaches only improved precision by arround 0.5%. The lower precision of *BSSM*'s approach in DO is due to the lower quality of the input alignment generated for the DO pair. The proposed strategies effectively handle this issue filtering out many erroneous mappings. *BSSM* finds 10 to 20 times more new mappings than other matchers. This was expected since the *BSSM* strategy only adds new mappings without computing any new score to filter out some of the candidates.

Table 5.7: Number of new mappings found by the strategies. their distribution according to the type of relationship and the precision of those new mappings *

OP	Strategy	NM	is_a	part_of	eq	DSD	neg	P(%)	NME
DO	Stem0.6Hyb0.6_0.75	98	57	0	1	1	2	95.0	61
	Stem0.6Per0.6	98	57	0	1	1	2	95.0	61
	Wor0.6Hyb0.6_0.75	73	47	0	1	1	2	94.0	51
	Wor0.6Per0.6	73	47	0	1	1	2	94.0	51
	Str0.6Hyb0.6_0.7	434	59	0	1	3	10	84.3	73
	Str0.6Hyb0.6_0.75	409	53	0	1	3	10	82.8	67
	BSSM0.6Per0.6	4465	85	0	0	2	27	75.9	114
HM	Wor0.6Hyb0.6_0.75	208	60	4	2	0	5	90.1	71
	Wor0.6Per0.6	208	60	4	2	0	5	90.1	71
	Stem0.6Hyb0.6_0.75	393	77	4	3	1	6	90.0	91
	Stem0.6Per0.6	393	77	4	3	1	6	90.0	91
	BSSM0.6Per0.6	5226	118	8	1	0	14	89.4	141
	Str0.6Hyb0.6_0.7	1002	76	9	3	1	10	86.7	99
	Str0.6Hyb0.6_0.75	929	67	8	3	1	9	86.2	88

* OP: Ontology Pair, DO: DOID-ORDO, HM: HP-MP; BSSM: Basic Semantic Subsumption Matcher, Wor: SubsWord, Stem: SubsStemmer, Str: SubsString; Per: Permissive Selector, Hyb: Hybrid Selector; DO: DOID-ORDO, HM: HP-MP; NP: New Mappings, eq: equivalences, DSD: Different Subsumption Direction, neg: negative; P: Precision, NME: number of mappings evaluated.

OAEI has a reference alignment but the major evaluation is through silver standards that assume a mapping correct if found by two or more systems. The other evaluation is to randomly select 30 of the unique mappings and manually evaluate them. Therefore this manual evaluation has complemented existing standards by showing that the precision of the new mappings can variate between 75.9% and 95.00%.

Chapter 6

Conclusion

This dissertation focused in the use of semantic similarity as part of ontology alignment algorithms in the biomedical domain. The underlying hypothesis was that semantic similarity could function as an extension technique to find novel mappings based on an anchor alignment. Two major contribution were achieved: (i) the development of a general algorithm for equivalence matching, which is parameterizable to conform to different types of ontologies; and (ii) the development of an algorithm for subsumption matching.

Given the fact that the equivalence matcher extension algorithm is parameterizable, its evaluation was done by grouping the parameters in three dimensions: structure, semantic similarity and weighting mechanism.

Since the evaluation was done by using distinct ontology pairs, it was possible to conclude that there is no single parameterization that produces a more accurate alignment than the others; therefore, I grouped the parameters into the three dimensions mentioned above and reported the average of the performance statistics of the top 20 best results for each possible parameter value in these dimensions. This allowed me to study whether a given parameter value is robust, i.e. whether there is one combination of parameters that repeatedly performs better than the others.

The main result of this evaluation strategy was that the best parameter combination depends on the characteristics of the ontologies being matched. In particular, the size and granularity of the ontologies have a strong impact on the best parameterization of the matching algorithm.

When compared to the baseline (a simple matcher using lexical information), the results show that there is an average increase of 0.2% in the F-measure. Notice that this results from the average of the top 20 strategies, which means that the correct parameterization can lead to a larger increase in this statistic.

When compared to the results obtained with the full AML pipeline, no significant improvements were achieved. This happens because the AML algorithm is already fine-tuned to work with the biomedical ontologies of OAEI. AML is a top performing ontology matching system meaning that even the smallest improvements are difficult to obtain. The comparison of the new algorithms to another top performing system (LogMap family) shows that even when using an input alignment that lacks Background Knowledge, the proposed approach (*Baseline*) extended by the algorithm can surpass LogMap (which uses Background Knowledge) by 11.3% on the F-measure results.

The equivalence matcher algorithm can only search the area that the input alignments provide. The recall of the found correct mappings in comparison to the possible ones in the radius defined are between 66.7% and 100%. Therefore the algorithm is able to finding correct mappings in the neighborhood it searches.

The subsumption matcher was evaluated through a manual evaluation that showed that between 75.9% and 95% of the new mappings found were correctly labeled. In particular, the combination of the semantic approach with lexical similarities that explore the implicit semantics of biomedical ontology terms proved to be successful.

The main conclusion of this work is that semantic similarity can contribute to ontology matching in the extension of existing alignments. The proposed algorithm for equivalence mappings is highly parameterizable and while it is not a single but a combination of parameters that works in all ontology pairs, this work hints at some possible good parameter selection for medium and small ontology pairs, namely Ancestors with radius 3, Seco with Resnik, and SSc_A/FSS_B. The subsumption matcher, although it was not possible to validated with a full reference, alignment manual evaluation results show that it is effective in finding actual hierarchical relationships between concepts from two ontologies.

The more interesting result of this dissertation is that using semantic similarity to extend high precision anchor alignments can be a valid option for ontology

matching in domains where Background Knowledge is unavailable or difficult to explore. Possible avenues for future research could include a more in depth evaluation of the comparison between semantic similarity based matching and Background Knowledge matching, and an evaluation taking as input anchor alignments generated by other systems. In fact, the lexical similarity used by AML, although simple in its genesis, takes into consideration different weights for different kinds of labels (e.g., main label, synonym, narrow synonym) (Pesquita *et al.*, 2013b).This alone has been shown to have a decisive impact in the biomedical ontology matching field (Faria *et al.*, 2018). It would then be interesting to explore whether semantic similarity based matching could achieve good performance using non-weighted lexical similarity and thus function as a more generic approach in this sense.

References

Achichi, M., Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G., Fundulaki, I., Harrow, I., Ivanova, V. *et al.* (2016). Results of the ontology alignment evaluation initiative 2016. In *CEUR workshop proceedings*, vol. 1766, 73–129, RWTH. 29, 31

Alexander, C.Y. (2006). Methods in biomedical ontology. *Journal of biomedical informatics*, **39**, 252–266. 10

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**, 25–29. 8

Bock, J. & Hettenhausen, J. (2012). Discrete particle swarm optimisation for ontology alignment. *Information Sciences*, **192**, 152–173. 16

Cross, V. & Hu, X. (2011). Using semantic similarity in ontology alignment. In *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*, 61–72, CEUR-WS. org. 15

Cruz, I.F., Palandri Antonelli, F. & Stroe, C. (2009). AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *PVLDB*, **2**, 1586–1589. 16

David, J., Guillet, F. & Briand, H. (2006). Matching directories and OWL ontologies with AROMA. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, 830–831, ACM. 16

REFERENCES

Diallo, G. (2014). An effective method of large scale ontology matching. *Journal of biomedical semantics*, **5**, 44. 16

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, **121**, 279. 30

Euzenat, J., Shvaiko, P. *et al.* (2007). *Ontology matching*, vol. 18. Springer. 6, 12, 13

Fahad, M., Qadir, M.A., Noshairwan, M.W. & Iftakhir, N. (2007). DKP-OM: A semantic based ontology merger. In *Proc. 3rd International Conference I-Semantics*, 313–322. 16

Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F. & Couto, F.M. (2013a). The agreementmakerlight ontology matching system. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 527–541, Springer. 32

Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F. & Couto, F.M. (2013b). The AgreementMakerLight Ontology Matching System. In *OTM Conferences - ODBASE*, 527–541. 16

Faria, D., Pesquita, C., Mott, I., Martins, C. & Couto, F.M. (2018). Tackling the challenges of matching biomedical ontologies. *Journal of Biomedical Semantics*. xi, 4, 8, 15, 55

Garcia, M.A.R., Gkoutos, G.V., Schofield, P.N. & Hoehndorf, R. (2016). Integrating phenotype ontologies with phenomenet. *Ontology Matching*, 201. 15, 16

Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J. & Parsia, B. (2011). The national cancer institute's thesaurus and ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, **1**. 30

Gruber, T. (2008). Ontology. *Entry in the Encyclopedia of Database Systems*. 5

Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, **43**, 907–928. 5

Gulić, M., Vrdoljak, B. & Banek, M. (2016). Cromatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment. *Web Semantics: Science, Services and Agents on the World Wide Web*, **41**, 50–71. 16

Hamdi, F., Reynaud, C. & Safar, B. (2010). Pattern-based mapping refinement. *Knowledge Engineering and Management by the Masses*, 1–15. 16

Harispe, S., Ranwez, S., Janaqi, S. & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, **8**, 1–254. 34

Hayamizu, T.F., Mangan, M., Corradi, J.P., Kadin, J.A. & Ringwald, M. (2005). The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome biology*, **6**, R29. 2, 30

Hertling, S. & Paulheim, H. (2012). Wikimatch: using wikipedia for ontology matching. In *Proceedings of the 7th International Conference on Ontology Matching-Volume 946*, 37–48, CEUR-WS. org. 16

Jean-Mary, Y.R., Shironoshita, E.P. & Kabuka, M.R. (2009). Ontology matching with semantic verification. *Web Semantics: Science, Services and Agents on the World Wide Web*, **7**, 235–251. 15, 16

Jimenez-Ruiz, E., Grau, B.C., Horrocks, I. & Berlanga, R. (2010). Towards a umls-based silver standard for matching biomedical ontologies. In *Proceedings of the 5th International Conference on Ontology Matching-Volume 689*, 220–221, CEUR-WS. org. 31

Jiménez-Ruiz, E., Grau, B.C. & Zhou, Y. (2011). Logmap 2.0: towards logic-based, scalable and interactive ontology matching. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, 45–46, ACM. 14, 16

REFERENCES

Jiménez-Ruiz, E., Grau, B.C. & Cross, V.V. (2016). Logmap family participation in the oaei 2016. In *OM@ ISWC*, 185–189. 14, 15

Kang, J. & Naughton, J.F. (2003). On schema matching with opaque column names and data values. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 205–216, ACM. 13

Kirsten, T., Gross, A., Hartung, M. & Rahm, E. (2011). GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *Journal of biomedical semantics*, **2**, 6. 16

Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J. *et al.* (2013). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, **42**, D966–D974. 8, 30

Kotis, K., Katasonov, A. & Leino, J. (2012). Aligning smart and control entities in the iot. *Internet of Things, Smart Spaces, and Next Generation Networking*, 39–50. 16

Lambrix, P. & Tan, H. (2006). Sambo—a system for aligning and merging biomedical ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web*, **4**, 196–206. 16

Li, W. (2015). Combining sum-product network and noisy-or model for ontology matching. In *OM*, 35–39. 16

Megdiche, I., Teste, O. & Trojahn, C. (2016). An extensible linear approach for holistic ontology matching. In *International Semantic Web Conference*, 393–410, Springer. 16

Meng, L., Gu, J. & Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in wordnet. *International Journal of Grid and Distributed Computing*, **5**, 81–94. 10

Müller, A.C. & Paulheim, H. (2015). Towards combining ontology matchers via anomaly detection. In *OM*, 40–44. 16

Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. & Haendel, M.A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome biology*, **13**, R5. 15

Nagy, M., Vargas-Vera, M. & Motta, E. (2007). DSSim: managing uncertainty on the semantic web. In *Proceedings of the 2nd International Conference on Ontology Matching-Volume 304*, 160–169, CEUR-WS. org. 16

Ngo, D., Bellahsene, Z. & Coletta, R. (2012). Yam++-a combination of graph matching and machine learning approach to ontology alignment task. *Journal of Web Semantics*, **16**. 16

Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O. & Couto, F.M. (2008). Metrics for go based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, **9**, S4. 11

Pesquita, C., Faria, D., Falcao, A.O., Lord, P. & Couto, F.M. (2009). Semantic similarity in biomedical ontologies. *PLoS computational biology*, **5**, e1000443. 9

Pesquita, C., Stroe, C., Cruz, I.F. & Couto, F.M. (2010). BLOOMS on AgreementMaker: Results for OAEI 2010. In *Proceedings of the 5th International Conference on Ontology Matching-Volume 689*, 134–141, CEUR-WS. org. 16

Pesquita, C., Faria, D., Santos, E. & Couto, F.M. (2013a). To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *Proceedings of the 8th International Conference on Ontology Matching-Volume 1111*, 13–24, CEUR-WS. org. 25

Pesquita, C., Faria, D., Stroe, C., Santos, E., Cruz, I.F. & Couto, F.M. (2013b). What's in a 'nym'? synonyms in biomedical ontology matching. In *International Semantic Web Conference*, 526–541, Springer. 55

REFERENCES

Quesada-Martínez, M., Fernández-Breis, J.T. & Stevens, R. (2013). Lexical characterization and analysis of the bioportal ontologies. In *Conference on Artificial Intelligence in Medicine in Europe*, 206–215, Springer. 28

Rahm, E. & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal*, **10**, 334–350. 13

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Journal of Artificial Intelligence Research*, :95–13. 10, 11

Reul, Q. & Pan, J.Z. (2010). KOSIMap: Use of description logic reasoning to align heterogeneous ontologies. In *23rd International Workshop on Description Logics DL2010*, 489. 16

Rosse, C. & Mejino, J.L. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, **36**, 478–500. 2

Rosse, C. & Mejino Jr, J.L. (2008). The foundational model of anatomy ontology. In *Anatomy Ontologies for Bioinformatics*, 59–117, Springer. 29

Rubin, D.L., Shah, N.H. & Noy, N.F. (2007). Biomedical ontologies: a functional perspective. *Briefings in bioinformatics*, **9**, 75–90. 9

Sánchez, D., Batet, M. & Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, **24**, 297–303. 11

Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G. & Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, **40**, D940–D946. 30

Seco, N., Veale, T. & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European conference on artificial intelligence*, 1089–1090, IOS Press. 10

Seddiqui, M.H. & Aono, M. (2009). An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size. *Web Semantics: Science, Services and Agents on the World Wide Web*, **7**, 344–356. 16

Smith, C.L. & Eppig, J.T. (2012). The mammalian phenotype ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian genome*, **23**, 653–668. 30

Stoilos, G., Stamou, G. & Kollias, S. (2005). A string metric for ontology alignment. *The Semantic Web–ISWC 2005*, 624–637. 27

Szwabe, A., Misiorek, P. & Walkowiak, P. (2012). Tensor-based relational learning for ontology matching. In *KES*, 509–518. 16

Tigrine, A.N., Bellahsene, Z. & Todorov, K. (2015). Light-weight cross-lingual ontology matching with LYAM++. In *ODBASE: Ontologies, DataBases, and Applications of Semantics*, 9415, 527–544. 16

Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P.N., Parkinson, H. & Rath, A. (2014). Ordo: An ontology connecting rare disease, epidemiology and genetic data. In *Proceedings of ISMB*. 30

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al.* (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, **36**, D13–D21. 37

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T. & Musen, M.A. (2011). Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, **39**, W541–W545. 9, 36

Zghal, S., Kachroudi, M., Yahia, S.B. & Nguifo, E.M. (2009). OACAS-Ontologies Alignment using Composition and Aggregation of Similarities. In *KEOD*, 233–238. 16

Zhao, M. & Zhang, S. (2016). Identifying and validating ontology mappings by formal concept analysis. In *OM@ ISWC*, 61–72. 16